USING DATA MINING TO DETERMINE THE IMPACT CONTINUITY OF CARE HAS ON THE AIR FORCE'S HEALTHCARE SYSTEM

THESIS
DECEMBER 2014

David F. Wade, Captain, USAF

AFIT-ENV-MS-14-D-44

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENV-MS-14-D-44

USING DATA MINING TO DETERMINE THE IMPACT CONTINUITY OF CARE HAS ON THE AIR FORCE'S HEALTHCARE SYSTEM

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Systems Engineering

David F. Wade, BS

Captain, USAF

December 2014

AFIT-ENV-MS-14-D-44

USING DATA MINING TO DETERMINE THE IMPACT CONTINUITY OF CARE HAS ON THE AIR FORCE'S HEALTHCARE SYSTEM

David F. Wade, BS

Captain, USAF

Approved:

//SIGNED//	12 Dec 2014
Christina Rusnock, Major, USAF, PhD (Chair)	Date
//SIGNED//	12 Dec 2014
Kyle Oyama, Lieutenant Colonel, USAF, PhD (Member)	Date
//SIGNED//	12 Dec 2014
Brian Stone, Major, USAF, PhD (Member)	Date

Abstract

Department of Defense (DoD) healthcare is one of the largest contributors to the DoD budget. In recent years, the cost of the DoD healthcare system has risen at an exponential rate. Much research has been conducted on the impacts that continuity of care has on both improving the quality of patient care and on reducing healthcare costs in the private sector. The DoD has attempted to take a similar approach with regards to healthcare continuity as a means to reduce healthcare costs. This research investigates whether continuity of care influences costs and a military member's availability to perform duties. Specifically, this research examines Air Force fliers with musculoskeletal injuries. Linear and logistic regression techniques are utilized to interpret the relationship continuity of care has on both patient availability and costs. The study does not identify any relationship between continuity of care with costs and patient availability. These findings suggest the need for further research as to whether these findings regarding continuity of care extend beyond musculoskeletal injuries within the DoD healthcare system, as well as evaluating other potential outcomes for continuity of care. Research should also be conducted to determine other factors influencing costs and patient availability.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Major Christina Rusnock, for her guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would, also, like to thank my sponsor, Lieutenant Colonel Anthony Tvaryanas and my data broker, Ms. Genny Maupin, both from the 711th Human Performance Wing at Wright Patterson AFB, OH for their subject matter expertise and assistance throughout this research.

David F. Wade

Table of Contents

Page

List of Figures

Page

List of Tables

Page

USING-DATA MINING TO DETERMINE THE IMPACT CONTINUITYOF CARE HAS ON THE AIR FORCE'S HEALTHCARE SYSTEM

I. Introduction

Background

The United States (U.S.) has seen rapid growth in healthcare costs; this rapid growth poses a major threat to the country's economic security and the security of its citizens (Schieber, et al., 2009). Though healthcare costs in the U.S. have grown faster than other similarly advanced and developed countries, the quality has not grown at a comparable rate. The U.S. is experiencing lower life expectancy and higher infant mortalities than other countries with lower healthcare costs (Farrell, 2008). The rate at which healthcare costs in the U.S. are growing is unsustainable (Mitchell, 2013). But why is the cost of healthcare within the U.S. increasing so rapidly? The rapid increase in healthcare cost can be attributed to many different factors. Technology is the most common factor attributed to healthcare cost growth. New technology is estimated to account for between 38 and 65 percent of cost growth in U.S. healthcare system (Schieber, et al., 2009). Administrative costs also account for a great portion of healthcare cost growth as well, with an average growth rate of 7 percent between 1995 and 2005 (Farrell, 2008). Lastly, price insensitivity of patients coupled with healthcare providers' fear of malpractice lawsuits drive the providers to implement the most costly treatment options rather than lower cost treatment options (Farrell, 2008). Though many other factors have been noted for contributing to healthcare cost growth, these are some of the major drivers to the unsustainable growth of healthcare costs within the U.S.

Though all of these factors are noted as having adversely affected healthcare in the private sector, government programs are not exempt from some of these same issues. Specifically, healthcare costs in the Department of Defense (DoD) are also on the rise at a rapid pace. Some of the factors affecting DoD healthcare costs include expanded benefits and increased usage of healthcare benefits by eligible beneficiaries. DoD healthcare accounts for nearly one tenth of the total DoD budget (Harrison, 2010). In an environment of economic conservatism, finding ways to decrease costs for government programs is highly desirable, particularly DoD healthcare. In accordance with the Pareto Principle, it is assumed that 20% of patients within the healthcare system consume 80% of the resources contributing to the higher majority of healthcare cost (Weinberg, 2009). Identifying the hypothesized high cost group within the DoD that accounts for a majority of costs and determining trending characteristics of this population could be beneficial in forecasting ways of preventing common healthcare issues and can ultimately reduce costs. One potential solution to reduce healthcare costs, while improving quality, is to increase continuity of care. Evidence suggests that there is an association between higher continuity of care and lower healthcare costs (Kristjansson, et al., 2013; Mainous & Gill, 1998).

Continuity of care is the continual process of care by the same healthcare provider and its patients over time. Over time the healthcare provider can establish rapport with the patient, identify patient trends, minimize repeat diagnostic testing, and provide more effective, higher quality care. Substantial literature has been developed on healthcare continuity in the private sector. The DoD implements continuity of care by requiring compliance with the standards and guidelines of the Patient Centered Medical Home

(PCMH) model requiring the continuity of medical record information at all times and monitoring the percentage of patient visits with a selected clinician or team (PCMH, 2014). Unfortunately, unique attributes in the DoD, such as frequent deployments and relocations, make healthcare continuity more difficult than that of the private sector.

Problem Statement

The 711th Human Performance Wing (HPW) at Wright Patterson Air Force Base in Ohio has vested interest in data analysis that can identify ways to reduce healthcare costs within the Air Force. The current Air Force healthcare model specifies that patients should meet with their primary care manger (PCM), also known as primary care provider, for at least 90% of their appointments and should meet with a member of their PCM team for at least 70% of their appointments. Although the Air Force healthcare model accounts for continuity of care, no empirical analysis and evidence exists that validates its benefits. This research seeks to fill this gap by evaluating the impact continuity of care has on healthcare costs and the readiness of Air Force personnel. To effectively conduct this analysis, this research limits its evaluation to active duty fliers with musculoskeletal injuries (MSIs) due to the type of data available. For more discussion on the selection of this subpopulation see the Section Defining Cost Groups.

Research Objectives

Extensive review of the literature on healthcare analysis brings to light a gap within data analysis practices used within the private sector's healthcare system and the Air Force's healthcare system. The purpose of this thesis is to bridge that gap by using

similar data analysis techniques from the private sector, and implementing them in the Air Force healthcare system by tailoring to the unique characteristics of the Air Force.

The most applicable data mining techniques are used to analyze the data provided by the 711th HPW; the analysis identifies the portion of the active duty fliers with MSIs within the Air Force that accounts for the highest percentage of healthcare costs. The analysis also identifies which characteristics and diagnoses are predictive of costs across both low and high cost groups and how continuity of care impacts healthcare costs and patient availability.

Investigative Questions

A series of investigative questions were developed to guide the research. The subpopulation referred to below consist of the active duty fliers with MSIs selected for evaluation by this analysis.

1) What percentage of the subpopulation contributes to a majority of the healthcare costs?
2) What are the defining characteristics of the high cost group?
    a. Which personal characteristics (gender, age group, race group, fitness information) correlate to higher healthcare costs?
    b. Which organizational factors (military rank and career field) account for higher healthcare costs?
3) How does continuity of care impact healthcare costs? Does the impact differ for high vs. low cost populations?

4) How does continuity of care impact patient availability? Does the impact differ for high vs. low cost populations?

To test the hypothesis of the Pareto Principle, this analysis begins with determining whether or not there is a small portion of the subpopulation that contributes to the majority of healthcare costs. Once this is established, the high cost and low cost groups are analyzed separately for comparison to determine which personal and organization characteristics are prominent in each group and if there is evidence that certain characteristics are predictive of costs. The research also investigates the impact continuity of care has on healthcare costs and patient availability. Answering these questions provides beneficial insight into the current Air Force healthcare model and how to better implement continuity of care.

Methodology

Due to the wide and successful use of data mining techniques in the healthcare industry, these methods are used to analyze the Air Force's continuity of care healthcare data for fliers with MSIs. Specifically multivariate linear regression, logistic regression, and simple linear regression are used. The results of the data mining analysis reveal the specific common characteristics of the high cost group within the Air Force. These characteristics provide insight into the specific demographic and organizational factors that correlate to higher healthcare costs of Air Force personnel. The results also provide insight into the Air Force's current continuity of care model and demonstrate whether increased continuity of care is correlated with decreased costs and increased patient availability for fliers with MSIs.

Assumptions and Limitations

In order to perform analysis, certain assumptions and limitations are made regarding the data.

Assumptions

- Patient appointment costs only include costs incurred for services rendered; these costs do not include fixed costs.
- Assume the data to be accurate
- Assume the data to be complete

Limitations

- Unable to obtain data on duty location, deployment information, and the aircraft the patient is assigned to
- For privacy purposes, data is limited to:
    - Age groups; actual age is not included for privacy purposes
    - Rank groups as opposed to specific military rank title
    - Appointment year; actual dates not of each appointment are not included

Data Scoping and Handling

It is necessary to scope down the problem in order to create a more manageable dataset. To do this, assumptions are made for this research. First, the population is reduced to active duty Air Force fliers with musculoskeletal injuries (MSIs). The dataset is scoped down to Air Force fliers because the Air Force healthcare system accurately tracks patient availability through pilots' flying status codes. For non-fliers, "profiles"

are set for those members who are unavailable for duty. Profiles are commonly unreliable in determining a patient's actual availability. Profiles frequently expire before a patient has fully recovered, or are not updated in the system when a patient recovers earlier than anticipated. Using flying status codes for fliers allows a more accurate depiction of a patient's availability. MSIs are considered because they provide a wide range of costs due to the considerable flexibility in diagnoses, diagnostic methods, and treatments. Second, healthcare costs will include only costs incurred for services rendered; they will not include administrative or overhead costs, given these are costs not specific or influential to a patient's quality or continuity of care.

This thesis utilizes centralized medical databases maintained by the Air Forces Surgeon General (AF/SG6). All data are stripped of personal identifiers before analysis is performed. The data are housed on existing computers in the Human Systems Laboratory at the Air Force Institute of Technology at Wright Patterson Air Force base in Ohio. These computers require Common Access Card enabled access granted to government employees and contractors, with the data stored in limited permissions directories.

Preview

This chapter provides the motivation and importance for a need for further research of the Air Force's healthcare system. Chapter II gives a background on the literature that exists on data mining within private sector healthcare, military healthcare applications of data mining, and continuity of care within the private sector. Chapter III gives an overview of the methods and processes used to perform the analysis and answer

the investigative questions. Chapter IV presents the results of the analysis and how it is interpreted. Chapter V provides the key conclusions to be drawn from the research and offers recommendations on future research on the topic of healthcare within the Air Force.

II. Literature Review

Chapter Overview

This chapter examines the background and literature of data mining techniques in healthcare and the impacts of implementing healthcare continuity. Data mining has been evolving as a more robust way to analyze large datasets. With the development of electronic healthcare records, data mining is essential to progression and advancement within the medical community. The use of data mining in healthcare can provide insights into better treatment regimens and earlier detection and prediction of chronic illnesses. This chapter will review the literature by exploring the implementation of data mining techniques in different areas of the healthcare community. With the DoD having the most robust healthcare records system in the country (Dolfini-Reed & Jebo, 2000), this chapter will also review the applications of data mining within the DoD healthcare system.

It is hypothesized that continuity of care in a healthcare system decreases a patient's likelihood of future hospitalization and increases the quality of care experienced by the patient (Mainous & Gill, 1998). This chapter reviews the literature that exists on continuity of care and the impacts continuity of care has on patient quality and healthcare costs.

Data Mining

With increased technology, data is being collected and stored at a rapid pace. Data mining assists in managing and analyzing large datasets (Fayyad, Piatesky-Shapiro, & Smyth, 1996). Data mining, commonly referred to as the knowledge discovery of

databases (KDD), is a comprehensive term that describes a combination of statistical and computer science techniques to discover relationships and patterns within large databases (Srinivas, Kavihta, & Govrdhan, 2010). Medical databases have been increasing in size making traditional data analysis methods much more difficult. Data mining has evolved from these traditional analysis methods to create algorithms to extract patterns from data. There are a variety of data mining methods utilized across a variety of applications including marketing, investments, fraud detection, manufacturing, and healthcare (Fayyad, Piatesky-Shapiro, & Smyth, 1996).

Data Mining within Healthcare

Given the size of medical records and information, data mining is an essential tool to healthcare reform and the efficiency of medical processes. The conversion to electronic medical records over the years has created the ability to gather more healthcare data (Prather, et al., 1997). With the dramatic growth in the size of medical databases, manual data analysis is impractical (Fayyad, Piatesky-Shapiro, & Smyth, 1996). Because of this, data mining has become more popular and critical within the healthcare community.

Multiple data mining techniques are being utilized within the healthcare community, including factor analysis (Fayyad, Piatesky-Shapiro, & Smyth, 1996), multivariate analysis (Gilmer, et al., 2005; Reid, et al., 2009) univariate and multivariate logistic regression (Lv, et al., 2011; Kurth, Glynn, Gaziano, Berger, & Robins, 2006), and multivariate time series algorithms (Wong, 2004). Extensive research exists in a variety

of different areas of healthcare and data mining from detecting disease outbreaks to the implementation of patient-centered medical home model.

Benefits of Healthcare Data Mining

Data mining is used in healthcare to improve effectiveness of treatments, healthcare management, and healthcare quality (Koh & Tan, 2011). Effectiveness of treatment is a measure of the effectiveness of the actions taken to move a patient from an unhealthy state to a healthy state. These actions incorporate a wide range of treatment options including pharmaceutical prescriptions, laboratory procedures, and simple doctor visits. There are several ways in which data mining has been used to measure how effective a treatment is for an illness. Kincade (1998) analyzes how effective and cost efficient specific drug regimens were for patients of the same condition. Srinivas (2010) utilizes decision tree analysis to predict the potential for a patient to experience a heart attack based on patient characteristics. Data mining is useful in finding root causes for more effective treatment.

Healthcare management is the ability to better track chronic illness and manage the illnesses appropriately; successful healthcare management is known to reduce hospital admissions and claims (Koh & Tan, 2011). Data mining has been used to mitigate issues of resource usage, management of hospital resources, and predict inpatient length of stay (Sharma & Mansotra, 2014). Kincade (1998) does this by categorizing patients according to demographic and medical conditions to help determine high cost populations based on resource utilization and frequency of visits. Data mining

can help identify areas of risk and improvement and provide valuable information to make the healthcare management process more effective.

Quality of care is the patient's satisfaction with the services provided as well as the short and long-term impacts of these services. Schuerenberg (2003) utilizes decision tree analysis to improve the quality of healthcare from treatment, disease management, and cost management. Brannigan (1999) implements a study that uses data mining as a tool to regulate patient wait times and improve service to patients. Data mining has many uses to help improve quality of care to patients.

Continuity of Care

It is hypothesized that continuity of care in a healthcare system decreases a patient's likelihood of future hospitalization (Mainous & Gill, 1998), ultimately decreasing healthcare costs. Generally, a patient's primary care provider is the first point of contact in the healthcare system (Balasubramanian, Banerjee, Denton, Naessens, & Stahl, 2010). Thus, in a long-term physician-patient relationship, a knowledge based is accrued (Mainous & Gill, 1998). Primary care providers are responsible for preventive medicine, patient education, routine physical exams, and referring patients to medical specialties for specialized care (Balasubramanian, Banerjee, Denton, Naessens, & Stahl, 2010). It is believed that physician and patient continuity is fundamental to good primary healthcare and is effective in reducing healthcare cost (Weiss & Blustein, 1996).

Literature suggests additional benefits of implementing continuity in a healthcare system include decreases in the number of appointments a patient will need, the number

of laboratory tests needed, and the overall number of emergency room visits (Weiss 1996). A primary care provider that has a relationship with the patient may perform more cost effectively with respect to their diagnosis (De Maeseneer, De Prins, Gosset, & Heyerick, 2003); managing the number of appointments is crucial to improving quality and managing costs (Green, Savin, & Murray, 2007); the number of appointments needed can be minimized through increased continuity (Tantau, 2009).

Reports by the Institute for Healthcare Improvement show that 40% of emergency department cases occurred because patients could not see their primary care provider (Balasubramanian, Banerjee, Denton, Naessens, & Stahl, 2010). Patients who meet regularly with their primary care providers are generally more satisfied with the care provided, more likely to take medications properly, more likely to be properly diagnosed, and less likely to be hospitalized (Balasubramanian, Banerjee, Denton, Naessens, & Stahl, 2010). Studies show that continuity of care is effective in lowering emergency room use, hospitalization, and reducing the number of no-shows for appointments (Kristjansson, et al., 2013). Quality of care and patient satisfaction is also shown to increase with continuity (Bjorkelund, et al., 2013).

Much research exists that shows the impact of patients meeting with their primary care providers and its impact on costs. In a survey analysis by Weiss & Blustein (1996), the results show that patients with high provider continuity (10+ years) experienced substantially lower costs of care. This cost association was also seen in a study observing the Belgian healthcare system over a two year period which showed that patients who visited the same family physician had lower total costs for medical care (De Maeseneer, De Prins, Gosset, & Heyerick, 2003). It is also important not to discount the research that

shows the impact continuity of care has on the quality of care. As noted previously, increased quality of care can result in reduced appointments, ultimately reducing healthcare cost. Mainous & Gill (1998) find that high continuity of care decreased the likelihood of future hospitalization. Anderson et al (2012) found that medical continuity was more common among older patients, and higher continuity resulted in a lower probability of needing emergency care and lower total medical costs. These studies show that increasing continuity will in time increase healthcare quality, ultimately reducing healthcare costs.

While many studies investigate continuity of care in the private sector, limited research exists on continuity of care within the military healthcare system. Given the unique nature of the military healthcare system with the frequent movement and deployment of its healthcare providers and members, the impacts of continuity of care are expected differ in the military healthcare system compared to that of the private healthcare system. Additionally, while extensive research exists that investigates costs and factors that influence costs, limited research explores factors that influence patient availability. For the military healthcare system, it is important that patients have rapid recoveries in order to be ready for duty. This also differentiates the military healthcare system from the private healthcare system.

Data Mining Military Applications

Data mining is used in the DoD in multiple areas including incidence ratio analysis to determine the frequency of incidences of cancer within the US Air Force active duty population (Yamane, 2006), correlation analysis of military personnel to link

illness among Gulf War veterans (Bose & Mahapatra, 2001), and scoring model to determine if patients with diabetes would require readmission (Ramachandran, Erraguntla, Mayer, & Benjamin, 2007). While military data mining applications are varied, limited research has focused specifically on cost, patient availability, or continuity of care within the military healthcare system. This research begins to close that gap through an analysis of these factors for Air Force active duty pilots with musculoskeletal injuries.

Conclusion

The purpose of this literature review is to provide the background on the literature that exists on data mining within healthcare. The chapter defines and provides an overview of data mining. Next, this chapter discusses data mining and how it has been used in the healthcare field in the private sector and its benefits. Then, the chapter examines the literature on continuity of care and its benefits in the private sector. Last, this chapter explores data mining applications in the military.

III. Methodology

Chapter Overview

The purpose of this chapter is to describe the methods used to analyze the factors that contribute to healthcare cost and patient availability. The chapter first examines the process of data gathering, collection, and formatting to prepare data for analysis. Next, it explores the problem formulation needed to effectively answer the investigative questions. Then, the three phases of the analysis process are explained along with the details of each step of the analysis process and the investigative questions that are being answered at each step. Lastly, the chapter summarizes the information covered.

Scoping of Data

Healthcare data are obtained from the Air Force's CarePoint site. The subpopulation chosen for this study is active duty fliers whose diagnose results in them being off of flying status due to musculoskeletal injuries (MSIs). Active duty fliers are chosen because of the accurate records kept on whether a patient is available for duty via their flying status; this allows a more accurate way of tracking patient availability than is possible for non-flying military personnel.

MSIs are chosen as the diagnosis of choice based on their frequency amongst fliers due to the strenuous activity associated with flying (Tvaryanas, 2014). In addition, MSIs provide a variety of different diagnoses types from less sever diagnoses such as back pain and joint pain, to more sever diagnoses such as bone disease and injuries of the spine. MSIs also provide a wide range of tools and procedures used to diagnosis and treat them (Tvaryanas, 2014). Since MSI diagnoses, diagnostics, and treatments are so

diverse, this set of conditions allows the results of the continuity of care analysis to be more generalizable to other non-MSI diagnoses.

Problem Formulation

The two dependent variables considered are healthcare cost and patient availability. Healthcare costs are the costs associated with providing care; these are costs associated with medical procedures, pharmaceutical prescriptions, and laboratory tests. Total costs are considered for each patient appointment over a five year period (July 2009-June 2014). Patient availability is calculated as the length of time a patient is off of flying status cumulatively from 2009. The independent variables considered are the personal and organizational factors listed in Table 1.

Table 1: Independent Variables

Gender
Male
Female
Age
Ages 19-29
Ages 30-39
Ages 40-49
Ages 50+
Race Group
Asian or Pacific Islander
Black, not Hispanic
Hispanic
Other/Unknown
White, not Hispanic
Military Rank / Level of Experience
Junior Enlisted
Senior Enlisted
Junior Officer
Senior Officer

Career Field
1. Pilot
2. Combat Systems Officer
3. Aicrew
4. Command and Control
5. Aircrew Protection
6. Flight Nurse
7. Aerospace Medicine Specialist
8. Aerospace Medical Service
9. Air Battle Manager / Special Tactics / Combat Rescue / Space Officers
Fitness Information
Height
Weight
Physical Fitness Test Run Score
Physical Fitness Test Score
Abdominal Circumference

Personal factors are those unique to the patient that the Air Force cannot control. The personal factors considered in this study are:

- Gender – measured as a binary variable, with "1" for male and "0" for female
- Age group- This consists of 4 dummy variables listed in Table 1. Each dummy variable is measured using a data field for each age group; measured with binary variable "1" if the patient is in age group and "0" if the patient is not
- Race group – This consists of 5 dummy variables listed in Table 1. Each dummy variable is measured using a data field for each race group; measured with binary variable "1" if the patient is in race group and "0" if the patient is not
- Fitness information- Includes height, weight, and abdominal circumference information. Test run score and physical fitness test score are measured on a 0 to 100 scale; 100 being the best score.
- Military rank – This consists of 4 dummy variables listed in Table 1. Each dummy variable is measured using a data field for each military rank group; measured with binary variable "1" if the patient is in military rank group and "0" if the patient is not. The ranks included in each rank group are listed below:
    - Junior Enlisted: Airman Basic, Airman, Airman First Class, Senior Airman
    - Senior Enlisted: Staff Sergeant, Technical Sergeant, Master Sergeant, Senior Master Sergeant, Chief Master Sergeant
    - Junior Officer: Second Lieutenant, First Lieutenant, Captain
    - Senior Officer: Major, Lieutenant Colonel, Colonel, General Officers

- Career field – This consists of 9 dummy variables listed in Table 1. Each dummy variable is measured using a data field for each career field; measured with binary variable "1" if the patient is in the career field and "0" if the patient is not

Investigative Questions

1) What percentage of the population contributes to a majority of the healthcare costs?
2) What are the defining characteristics of the high cost group?
    a. Which personal characteristics (gender, age group, race group, fitness information) correlate to higher healthcare costs?
    b. Which organizational factors (military rank and career field) account for higher healthcare costs?
3) How does continuity of care impact healthcare costs? Does the impact differ for high vs. low cost populations?
4) How does continuity of care impact patient availability? Does the impact differently for high vs. low cost populations?

First, it is important to begin the analysis identifying the high cost group and the defining characteristics of both low and high cost groups. This helps target specific groups in which improvements to healthcare costs could be most effective. Next, understanding the impact continuity of care has on patient appointment cost is important to help manage rising healthcare costs in the Air Force. Lastly, patient availability is essential in the Air Force's healthcare system because Air Force members need to be ready to deploy and support the Air Force's mission. Understanding how continuity of

care impacts patient availability is important to establish effective policies and requirements on continuity of care. Answering these investigative questions will provide beneficial insight into the current Air Force healthcare system and its effectiveness.

Methodology Phases

To organize the research process, the method is divided into three separate phases. These three phases answer the specific investigative questions where the appropriate analysis is required.

1. Data Collection
2. Defining Cost Groups
3. Regression Analysis

Data Collection

Data are collected from multiple sources: The Aviation Safety Information Management System (ASIMS), the Air Force Military Personnel Database (mil_pers), the Air Force Fitness Management System (AFFMS), the Cardiac Risk Management database (CRAM), and the Comprehensive Ambulatory/Professional Encounter Record (CAPER) database. This study has an approved Institutional Review Board (IRB) review and Health Insurance Portability and Accountability Act (HIPAA) waiver (Appendix A: IRB Approval Letters). A data broker removed personally identifiable information from the data prior to this analysis. The data obtained from the above databases is detailed below:

ASIMS: Contains information regarding duty, mobility, and flying status (patient availability).

Mil_pers: Contains personal and organizational factors that include the career field, gender, military rank, and age group data fields.

AFFMS: Contains data reflecting results of patients bi-annual physical fitness assessment; the assessment measures cardiac ability through a 1.5 mile run, number of push-ups and sit ups completed in one minute, body mass index (BMI), height, and weight.

CAPER: Contains detailed information regarding patient medical appointments. This database gives appointment costs information that include procedural, pharmaceutical, and laboratorial. It also includes details that show the diagnosis type, continuity of care information, and the year in which the patient was seen.

The data analyzed are for Air Force active duty fliers who are off of flying status due to an MSI diagnosis as of July of 2009; these data cover a five year period of patient appointment history from July 2009 to June 2014. Thus, active duty fliers with MSIs in July of 2009 are defined as the subpopulation for which analysis is conducted. Upon collection of the data, it is important to format the data to get it in a form usable to be analyzed to answer the investigative questions. The data are formatted in the following manner:

Data Assumptions

- Continuity of care only exists if a patient has more than one appointment; patients with one appointment are removed from the dataset.

- Patient availability is defined as number of days a patient is available to fly; since data is limited to flying status year, as opposed to flying status date, availability is looked at cumulatively starting with 2009 and will end at 2012. Patient availability is not calculated beyond 2012 because all patients had returned to flying status or had separated from the Air Force beyond 2012.

Data Formatting

- Medical appointments without at least one MSI diagnosis were removed.
- 61% of patients have appointments with missing provider IDs; to account for this, analysis is performed using two different scenarios:
    - Best case scenario: All blank provider ID entries appointments are assumed to be appointments with the same provider
    - Worst case scenario: All blank provider ID entries are assumed to be appointments with different providers
- MSI diagnoses were broken into four types:
    - Arthropathies – Diseases of the joints / joint inflammation
    - Dorsopathies – Spinal disease / injuries of the back
    - Rheumatism – Pain associated with joints and connective tissues (back pain, neck pain and osteoarthritis)
    - Osteopathies, chondropathies, and acquired musculoskeletal deformities – Diseases associated with bones or cartilage

Defining Cost Groups

Data mining is vital to understand the issues related to fliers and the specific organizational and personnel factors that contribute to healthcare cost. To begin the analysis, certain cost groups are identified by identifying the top percentages of the highest cost patients and calculating the percentage of total costs these patient's account for. Identifying the different cost groups allows the ability to analyze the data in smaller subsets that are more similar to rid the influence of results by more dominant groups. To establish the cost groups, each patient's total appointment costs are aggregated over the five year period. Once the patient's costs are calculated, patients are sorted in order by their total appointment costs. Potential cost groups are identified by the percentage contribution to total costs; the cost groups considered are the top 5%, 10%, 15%, 20%, 25%, and 30%. The break out that comes closest to the 80% hypothesized by the Pareto Principle is selected as the high cost group. The results of this cost group identification answers question 1 regarding identifying the percentage of the population that contributes to the preponderance of the healthcare costs.

Regression Analysis

Once the cost groups have been identified, analysis of variance (ANOVA) tests are performed on the two cost groups separately to test for attributes that are predictive of costs. Minitab (Version 15) is used to develop the initial ANOVA tables. Next, multivariate regression is used to quantify the impacts that personal and organizational factors have on healthcare costs. For the multivariate regression, the variables gender, age, race, rank, and career field are used as independent variables to the response

variable, cost. P-values from simple linear regressions are evaluated as a screening experiment, using a threshold of 0.05 to determine if a variable is predictive of costs. The characteristics identified with p-values less than 0.05 are included in the multivariate regression as predictive of cost within that given cost group. Logistic regression is performed to determine which characteristics are predictive in determining which cost group a patient belongs to. The results of these ANOVAs and regression analyses answer question 2 regarding identifying the cost groups and the defining characteristics of those cost groups.

Separate simple linear regression is performed on the independent variable, continuity of care, against the response variables, patient appointment cost and patient availability. Continuity of care is defined as the percentage of times the patient meets with their designated primary care manager for an illness whereas patient availability is defined as the number of days a patient was available to fly cumulatively since 2009. P-values for each simple linear regression equation are evaluated; cases in which the p-value are less than or equal to 0.05 are considered statistically significant.

The simple linear regression analysis for continuity of care against patient appointment cost is tested separately for each cost group, each specific diagnoses type, and scenario type. The cost groups will consist of the low cost group, high cost group, and all patients combined into a single group. The diagnosis types are arthropathies, dorsopathies, rheumatism, osteopathies, and all patient diagnoses to include both MSI and non-MSI diagnoses (MSI patients may have non-MSI diagnoses in the same appointment as an MSI diagnosis). The different scenarios are the best case (where blank provider IDs are considered the same provider) and worst case scenarios (where blank

provider IDs are considered to be different providers). Given that there are 3 cost groups, 5 different diagnosis types, and 2 different scenarios, a total of 30 simple linear regression graphs and equations are generated. This regression analysis answers question 3 regarding whether continuity of care impacts healthcare cost.

The simple linear regression analysis for continuity of care against patient availability is tested separately for each cost group, cumulative calendar year, and scenario type. The cost groups are also the low cost group and high cost group. The cumulative calendar years are 2009, 2010, 2011, and 2012. The scenario types are the best case and worst case scenarios. Given that there are 2 cost groups, 4 cumulative years, and 2 different scenarios, 16 total simple linear regression graphs and equations are generated for continuity of care vs. patient availability. This regression analysis answers question 4 which asks whether continuity of care impacts patient availability.

Conclusion

The purpose of this chapter is to provide the methodology for analyzing the impact continuity of care has on fliers with MSIs within the Air Force's healthcare system. The methods employed are multivariate linear regression, simple linear regression, and logistic regression. First, the characterization of the patients is determined for the different cost groups. Next the influences continuity of care has on healthcare cost and patient availability are evaluated and compared for both cost groups. These methods are sufficient in answering questions of whether continuity of care impacts fliers with MSIs within the Air Force's healthcare system.

IV. Analysis and Results

Chapter Overview

The purpose of this chapter is to present the results of the regression analysis completed to answer the investigative questions in regards to the impact continuity of care has on fliers with musculoskeletal injuries (MSIs). The results of this analysis help provide beneficial insights into the Air Force's current continuity of care model implemented in its healthcare system.

The investigative questions are divided into two categories: demographic characterization and continuity of care. The demographic characterization analysis is performed to determine which proportion of the population is high cost and which proportion is low cost. Multivariable regression is performed to determine if there are defining characteristics that make up each group; logistic regression is performed to evaluate if it can be determined which group a patient belongs to based upon known characteristics. For the continuity of care analysis, simple linear regression is performed to determine in which instances continuity of care has influence over patient appointment costs and patient availability.

Assumptions and Data Formatting

The dataset is comprised of patient appointment and characteristic information from July 2009 through June 2014. The dataset includes all patients that are off of flying status due to an MSI in July of 2009, and follows their medical appointment history through June 2014. To have the data in its clearest and most accurate representation, several assumptions are made and data formatting is performed.

Data Assumptions

- Continuity of care only exists if a patient has more than one appointment; patients with one appointment are removed from the database.
- Patient availability is defined as number of days a patient is available to fly in a given year. Because data is limited to flying status year, as opposed to flying status date, availability is looked at cumulatively starting with 2009 and will end at 2012. Patient availability is not calculated beyond 2012 because no patients from the July 2009 group are off of flying status due to an MSI beyond 2012.

Data Formatting

- Patients must have at least one medical appointment with an MSI diagnosis to be included.
- Numerous patients have appointments with missing provider IDs; to account for this, analysis is performed using two different scenarios:
    - Best case scenario: All blank provider ID entries appointments with the same provider
    - Worst case scenario: All blank provider ID entries are interpreted as appointments with different providers
- MSI diagnoses were broken into four types:
    - Arthropathies – Diseases of the joints / joint inflammation
    - Dorsopathies – Spinal disease / injuries of the back

- Rheumatism – Pain associated with joints and connective tissues (back pain, neck pain and osteoarthritis)
- Osteopathies, chondropathies, and acquired musculoskeletal deformities – Diseases associated with bones or cartilage

Demographic Characterization

Cost Profiles

Patient appointment costs are summed over the five year period for each patient for all appointments that include at least one MSI diagnosis, yielding a total cost per patient. Patients are sorted in order by their patient total costs; starting with the highest cost patients, the top 5% - 30% (in increments of 5%) are calculated along with their associated percentage of the subpopulation total cost. Table 2 contains the percentage of the subpopulation and their associated percentage of subpopulation total costs. This break out is used to identify the division of the subpopulation that best represents the 80-20 split hypothesized by the Pareto Principle. The top 30% of patients that make up 70% of the subpopulation total costs are chosen as the high cost group while the bottom 70% of patients that make up 30% of subpopulation total costs are the low cost group.

Table 2: Cost Profile Table

Percentage of People	Percentage of Costs
5%	27%
10%	40%
15%	50%
20%	58%
25%	65%
30%	70%

Organizational and Personal Characteristics

Table 3 displays summary characteristics for each of the cost groups. The compositions of each cost group in terms of personal and organizational characteristics are relatively similar for both profiles. The largest difference to note is the age category, which has a higher proportion of patients ages 30-39 in the high cost group and a higher proportion of patients ages 40-49 in the low cost group.

Table 3: Characterization Table

Gender	Low Cost	High Cost
Male	90%	85%
Female	10%	15%
Age	**Low Cost**	**High Cost**
Ages 19-29	15%	11%
Ages 30-39	38%	52%
Ages 40-49	43%	33%
Ages 50+	4%	4%
Race	**Low Cost**	**High Cost**
Asian or Pacific Islander	3%	4%
Black, not Hispanic	5%	6%
Hispanic	4%	6%
Other/Unknown	5%	4%
White, not Hispanic	82%	81%
Rank	**Low Cost**	**High Cost**
Junior Enlisted	11%	15%
Senior Enlisted	29%	33%
Junior Officer	15%	12%
Senior Officer	44%	39%
Career Field	**Low Cost**	**High Cost**
Pilot	36%	32%
Combat Systems Officer	11%	10%
Air Battle Manager / Special Tactics / Combat Rescue / Space Officers	6%	4%
Aicrew	34%	35%
Command and Control	6%	9%
Aircrew Protection	0%	2%
Flight Nurse	3%	4%
Aerospace Medicine Specialist	3%	2%
Aerospace and Operational Physiology	0%	0%
Aerospace Medical Service	1%	3%

For each cost group, analysis of variance tests are run for each categorical characteristic against the response variable patient appointment costs; analysis of variance (ANOVA) is used to determine if the mean cost values differ for each level of

the characteristic. If the mean of one level is different for that of another, then the values of that characteristic could be predictive of patient appointment costs. A p-value threshold of 0.05 is used for statistical significance. Figure 1 lists the resulted ANOVA tables for each characteristic for the low cost group. All p-values for each ANOVA table are ≥ 0.05, therefore fail to reject the null hypothesis that the mean values are the same for the levels of each characteristic. Thus, there are no characteristics that are predictive of patient appointment costs in the low cost group. Figure 2 shows the resulted ANOVA tables for each characteristic for the high cost group. All p-values for each ANOVA table are ≥ 0.05 therefore fail to reject the null hypothesis that the mean values are the same per characteristic. Thus there are no characteristics that are predictive of patient appointment costs in the high cost or low cost groups.

Additionally, it is important to note the large confidence intervals for the under-represented categories within each factor. For example, the variance in confidence intervals for the race groups excluding White, not Hispanic are much larger than that of the White, not Hispanic race group. That is, there is a large difference between the sample size of the majority categories and the minority categories. Thus, the lack of statistical difference between the means for these ANOVAS is at least partially due to the small sample sizes for some categories. There may actually be statistical differences between the mean costs of each category, but it would be essential to have increased sample sizes for the minority categories to validate this.

One-way ANOVA: Appt Costs versus Gender

Source	DF	SS	MS	F	P
Gender	1	504656	504656	0.14	0.704
Error	515	1799850376	3494855		
Total	516	1800355032			

S = 1869 R-Sq = 0.03% R-Sq(adj) = 0.00%

Level	N	Mean	StDev
Female	50	2643	1787
Male	467	2749	1878

Individual 95% CIs For Mean Based on Pooled StDev

```
(----------------*----------------)
(-----*----)
2400      2700      3000      3300
```

Pooled StDev = 1869

One-way ANOVA: Appt Costs versus Race Group

Source	DF	SS	MS	F	P
Race Group	4	29390561	7347640	2.12	0.077
Error	512	1770964472	3458915		
Total	516	1800355032			

S = 1860 R-Sq = 1.63% R-Sq(adj) = 0.86%

Level	N	Mean	StDev
Asian or Pacific Islande	18	1995	1472
Black, not Hispanic	26	2892	1852
Hispanic	20	2902	1707
Other/Unknown	28	1971	1666
White, not Hispanic	425	2803	1892

Individual 95% CIs For Mean Based on Pooled StDev

```
(-----------*-----------)
(---------*---------)
(----------*----------)
(---------*---------)
(-*--)
1400      2100      2800      3500
```

Pooled StDev = 1860

One-way ANOVA: Appt Costs versus Career Field

Source	DF	SS	MS	F	P
Career Field	7	30489420	4355631	1.25	0.272
Error	509	1769865612	3477143		
Total	516	1800355032			

S = 1865 R-Sq = 1.69% R-Sq(adj) = 0.34%

Level	N	Mean	StDev
Aerospace Medical Servic	4	3450	1651
Aerospace Medicine Speci	18	2788	2059
Aicrew	177	2870	1953
Air Battle Manager / Spe	32	3074	1962
Aircrew Protection	29	3161	1942
Combat Systems Officer	56	2786	1697
Flight Nurse	16	2868	1782
Pilot	185	2443	1786

Individual 95% CIs For Mean Based on Pooled StDev

```
(-----------------*-----------------)
(--------*--------)
(--*-)
(------*-----)
(------*-----)
(----*----)
(--------*--------)
(-*--)
2000      3000      4000      5000
```

Pooled StDev = 1865

One-way ANOVA: Appt Costs versus Age Group

Source	DF	SS	MS	F	P
Age Group	3	6794820	2264940	0.65	0.585
Error	513	1793560212	3496219		
Total	516	1800355032			

S = 1870 R-Sq = 0.38% R-Sq(adj) = 0.00%

Level	N	Mean	StDev
Ages 19-29	77	2659	1814
Ages 30-39	195	2875	1888
Ages 40-49	222	2674	1906
Ages 50+	23	2464	1484

Individual 95% CIs For Mean Based on Pooled StDev

```
(---------*----------)
(------*-----)
(-----*-----)
(------------------*------------------)
2000      2400      2800      3200
```

Pooled StDev = 1870

One-way ANOVA: Appt Costs versus Rank Group

Source	DF	SS	MS	F	P
Rank Group	3	20070981	6690327	1.93	0.124
Error	513	1780284051	3470339		
Total	516	1800355032			

S = 1863 R-Sq = 1.11% R-Sq(adj) = 0.54%

Level	N	Mean	StDev
Junior Enlisted	58	2660	1765
Junior Officer	79	2776	1785
Senior Enlisted	152	3021	2003
Senior Officer	228	2557	1816

Individual 95% CIs For Mean Based on Pooled StDev

```
(---------------*---------------)
(-------------*------------)
(---------*---------)
(-------*-------)
2400      2700      3000      3300
```

Pooled StDev = 1863

Figure 1: Low Cost Group ANOVA Tables

One-way ANOVA: Appt Costs versus Gender

```
Source   DF          SS       MS     F      P
Gender    1      504656   504656  0.14  0.704
Error   515  1799850376  3494855
Total   516  1800355032

S = 1869   R-Sq = 0.03%   R-Sq(adj) = 0.00%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level     N  Mean  StDev  ---------+---------+---------+---------+
Female   50  2643   1787  (---------------*---------------)
Male    467  2749   1878               (-----*----)
                          ---------+---------+---------+---------+
                                2400      2700      3000      3300

Pooled StDev = 1869
```

One-way ANOVA: Appt Costs versus Age Group

```
Source      DF          SS         MS     F      P
Age Group    3    451522854  150507618  1.32  0.267
Error      217  24657405024  113628595
Total      220  25108927878

S = 10660   R-Sq = 1.80%   R-Sq(adj) = 0.44%


Level        N   Mean  StDev
Ages 19-29   25  13917   8246
Ages 30-39  115  14368   9390
Ages 40-49   73  16744  12990
Ages 50+      8  19651  10855

           Individual 95% CIs For Mean Based on Pooled StDev
Level      -+---------+---------+---------+--------
Ages 19-29  (--------*-------)
Ages 30-39        (---*---)
Ages 40-49          (---*----)
Ages 50+         (--------------*--------------)
           -+---------+---------+---------+--------
          10000     15000     20000     25000

Pooled StDev = 10660
```

One-way ANOVA: Appt Costs versus Race Group

```
Source       DF          SS         MS     F      P
Race Group    4   176684757   44171189  0.38  0.821
Error       216  24932243121  115427051
Total       220  25108927878

S = 10744   R-Sq = 0.70%   R-Sq(adj) = 0.00%


Level                      N   Mean  StDev
Asian or Pacific Islande   8  12128   4128
Black, not Hispanic       13  14735   7699
Hispanic                  14  13438   7411
Other/Unknown              8  13849   5294
White, not Hispanic      178  15687  11447

                          Individual 95% CIs For Mean Based on Pooled StDev
Level                     -+---------+---------+---------+--------
Asian or Pacific Islande   (--------------*--------------)
Black, not Hispanic               (----------*----------)
Hispanic                        (----------*----------)
Other/Unknown                (--------------*--------------)
White, not Hispanic                    (--*---)
                          -+---------+---------+---------+--------
                          5000      10000     15000     20000

Pooled StDev = 10744
```

One-way ANOVA: Appt Costs versus Rank Group

```
Source       DF          SS         MS     F      P
Rank Group    3   258195169   86065056  0.75  0.523
Error       217  24850732709  114519506
Total       220  25108927878

S = 10701   R-Sq = 1.03%   R-Sq(adj) = 0.00%


                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level             N   Mean  StDev  ---------+---------+---------+---------+
Junior Enlisted  34  12832   7644  (-----------*-----------)
Junior Officer   27  16075  11269            (-------------*-------------)
Senior Enlisted  74  15397  10055               (-------*-------)
Senior Officer   86  15931  12000                  (------*-------)
                          ---------+---------+---------+---------+
                              12000     15000     18000     21000

Pooled StDev = 10701
```

One-way ANOVA: Appt Costs versus Career Field

```
Source         DF          SS         MS     F      P
Career Field    8   678627816   84828477  0.74  0.660
Error         212  24430300062  115237264
Total         220  25108927878

S = 10735   R-Sq = 2.70%   R-Sq(adj) = 0.00%


Level                      N   Mean  StDev
Aerospace Medical Servic   7  17378  12598
Aerospace Medicine Speci   4  24478  13823
Aicrew                    78  14514   9119
Air Battle Manager / Spe   8  19118  18410
Aircrew Protection        19  14244  10102
Combat Systems Officer    22  16061  15261
Command and Control        4  12828   8181
Flight Nurse               9  12256   4336
Pilot                     70  15565  10109

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level                     --------+---------+---------+---------+-
Aerospace Medical Servic         (--------*--------)
Aerospace Medicine Speci               (----------*----------)
Aicrew                             (--*-)
Air Battle Manager / Spe           (-------*--------)
Aircrew Protection               (----*----)
Combat Systems Officer             (---*----)
Command and Control       (-----------*----------)
Flight Nurse                  (-------*-------)
Pilot                              (--*-)
                          --------+---------+---------+---------+-
                              10000     20000     30000     40000

Pooled StDev = 10735
```

Figure 2: High Cost Group ANOVA Tables

An additional method, multivariate regression, is tested to determine which characteristics are predictive of a patient's appointment costs. In step-wise form, characteristics with p-values above 0.15 are removed until the characteristics left have p-values close to and below 0.05. Table 4 shows the results of the final multivariate regression on the low cost group. The characteristics that remain in this regression equation are Asian or Pacific Islander race group, Other/Unknown race group, and the pilot career field. With an adjusted r-squared value of 0.024, this model is not very predictive of costs, therefore there may be other characteristics not included in this dataset that explain the variability in patient appointment costs for the low cost group.

Table 4: Low Cost Group Multivariate Regression Table

The regression equation is

Patient_Appointment_Costs = 2972 - 802*Asian or Pacific Islander - 824*Other/Unknown - 450*Pilot

Predictor	**Coef**	**SE Coef**	**T**	**P**
Constant	2972.1	104.1	28.55	0
Asian or Pacific Islander	-802.1	443.3	-1.81	0.071
Other/Unknown	-824.4	359	-2.3	0.022
Pilot	-450.1	169.4	-2.66	0.008

S = 1845.53 R-Sq = 2.9% R-Sq(adj) = 2.4%

Alternatively, for the high cost group there is only one characteristic that meets the criteria for inclusion in the multivariate regression. With p-values of 0.083, the flight nurse career field characteristic is slightly above the value of 0.05 for statistical significance. Table 5 shows the regression equation and r-squared values for this equation. With an adjusted r-squared value of 0.009, this model is not predictive of costs, therefore there may be other characteristics not included in this dataset that explain the variability in patient appointment costs for the low cost group.

Table 5: High Cost Group Multivariate Regression Table

The regression equation is

Patient_Appointment_Costs = 15124 + 9354*Flight Nurse

Predictor	Coef	SE Coef	T	P
Constant	15123.8	721.9	20.95	0
Flight Nurse	9354	5366	1.74	0.083

S = 10634.1 R-Sq = 1.4% R-Sq(adj) = 0.9%

Logistic Regression

Logistic Regression is performed to identify which characteristics can be used to determine whether a patient will be in the high cost group or low cost group. The continuous characteristics, height, weight, abdominal circumference, and physical fitness run, and total score are used as they provide the most beneficial information and statistically significant p-values. The results are tested iteratively using binary logistic regression and outliers are removed by observing delta chi-square values. Table 6 shows the results of the logistic regression. As shown, all p-values are below 0.05 for all characteristics. All goodness of fit tests pass because all p-values are greater than 0.05. The odds ratios show that physical fitness test score has the strongest influence over whether a patient will end up in the high cost group; for each test value point increase the odds that the patient ends up in the high cost group increases by 15%. This is counter-intuitive because members that are more physically fit are expected to require less medical attention and therefore cost less. It is important to note that these results are only for the subpopulation, and are not indicative of all Air Force patients. A potential explanation of these results are members who perform better on the physical fitness test are more likely to engage in strenuous activity and therefore have potential to incur higher costs for MSI diagnoses. Additionally, for height, physical fitness run score, and

abdominal circumference have the strongest influence over whether a patient will end up in the low cost group; as these values increase by value of one, the odds that the patient ends up in the high cost group decreases by 15%, 12%, and 16% respectively. The result that increased abdominal circumference decreases the likelihood a patient will be in the high cost group is also unexpected. It is also important to note that abdominal circumference is not normalized for either height or gender, and thus require further inspection beyond increased size. With an odds ratio of 1.04, which is close to 1, weight minimally affects the likelihood a patient will end up in the high cost group.

Table 6: Logistic Regression Table

Logistic Regression Table							
Predictor	**Coef**	**SE Coef**	**Z**	**P**	**Odds Ratio**	**95% Lower**	**CI Upper**
Constant	2.46992	4.8301	0.51	0.609			
Height	-0.159321	0.07152	-2.23	0.026	0.85	0.74	0.98
Weight	0.0399098	0.011501	3.47	0.001	1.04	1.02	1.06
Physical Fitness Test Run Score	-0.125073	0.037619	-3.32	0.001	0.88	0.82	0.95
Physical Fitness Test Score	0.136253	0.044221	3.08	0.002	1.15	1.05	1.25
Abdominal Circumference	-0.169311	0.083556	-2.03	0.043	0.84	0.72	0.99
Log-Likelihood = -124.698							
	G = 22.603		DF = 5		P-Value = 0.000		
Goodness-of-Fit Tests							
Method	**Chi-Square**	**DF**	**P**				
Pearson	214.153	224	0.67				
Deviance	249.396	224	0.117				
Hosmer-Lemeshow	13.623	8	0.092				

Continuity of Care and Healthcare Costs

Continuity of care is defined as the percentage of times a patient meets with the same healthcare provider. Simple linear regressions are performed for each MSI diagnosis against continuity of care as well as for all diagnoses as a whole. Simple linear

regression is calculated using two scenarios: best case and worst case. The best case scenario assumes that appointments where the providers IDs are missing from the database are all the same provider. The worst case scenario assumes that appointments where the providers IDs are missing from the database are all different providers. The true value is estimated to fall between these two extremes.

Best Case Scenario

Table 7 displays the results of each linear regression for each combination of diagnosis and cost group. Highlighted in blue are the cases in which there p-values are $\leq$ 0.05. For the high cost group, the p-value is ≤ 0.05 for only the dorsopathy diagnosis whereas the low cost group has cases with p-values ≤ 0.05 for all diagnoses with the exception of osteopathies. When all patients care combined into a single group, p-values are ≤ 0.05 for all diagnosis types. Although p-values for each case highlighted in blue are below 0.05, R^2 values for each of these equations are very low. Thus, no true conclusions can be drawn about the true impact continuity of care has on patient appointment costs. Figure 3 shows the linear regression graphs for the statistically significant cases. Figure 4 and Figure 5 show both the residual versus fits plots and normal plots for residuals for all cases in which p-values are greater than or equal to 0.05. With the exception of arthropathy diagnoses for all patients, in all other residual versus fit plots there is a pattern that shows as continuity of care increases, the variability in patient appointment costs also increases. The small variability at the lowest levels of continuity of care has significant influence over the created regression lines with small p-values and small R^2 values. These violate the assumption that there is constant variance along the regression

line. Additionally, the normal plots for residuals clearly show that in all cases, the residuals are not normal about the linear regression equation. This violates the second regression assumption of normality. Thus these regression equations are not good models to determine the impact continuity of care has on patient appointment costs.

Table 7: Linear Regression Results Best Case Scenario

Best Case Scenario	High Cost Group	Low Cost Group	All Patients
	p = 0.407; R^2 = 0.009	**p = 0.019; R^2 = 0.0122**	**p = 0.00; R^2 = 0.0534**
Arthropathies	y = -358.75x + 7774.4	**y = -659.92x + 2193.2**	**y = 985.14x + 2553.4**
	p = 0.007; R^2 = 0.0756	**p = 0.00; R^2 = 0.1378**	**p = 0.00; R^2 = 0.0634**
Dorsopathies	**y = -8749.6x + 9739.7**	**y = -2187.5x + 3054.5**	**y = -5070.8x + 5878.9**
	p = 0.126; R^2 = 0.0106	**p = 0.003; R^2 = 0.0593**	**p = 0.016; R^2 = 0.0168**
Rheumatism	y = -3897.1x + 6964.6	**y = -984.87x + 1604.5**	**y = -3316.4x + 4684.7**
	p = 0.19; R^2 = 0.0182	p = 0.13; R^2 = 0.00	**p = 0.013; R^2 = 0.0336**
Osteopathies	y = -2031.7x + 4318	y = -30.78x + 847.96	**y = -2164.5x + 3485.3**
	p = 0.499; R^2 = 0.0021	**p = 0.00; R^2 = 0.0632**	**p = 0.00; R^2 = 0.0175**
All Diagnoses	y = -3387.3x + 14054	**y = -2477.1x + 4012.3**	**y = -6452x + 9503.3**

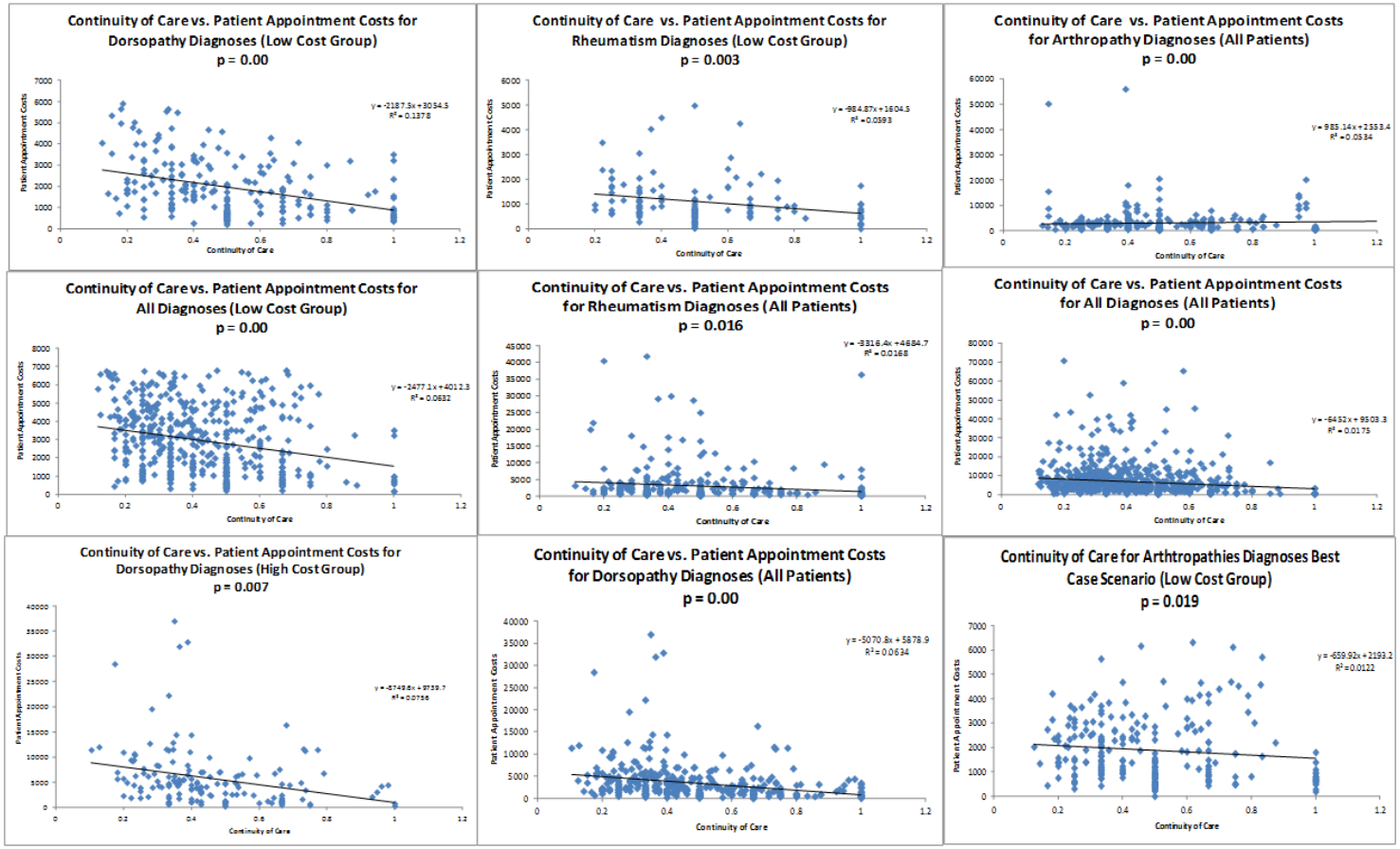

Figure 3: Best Case Scenario Regression Graphs

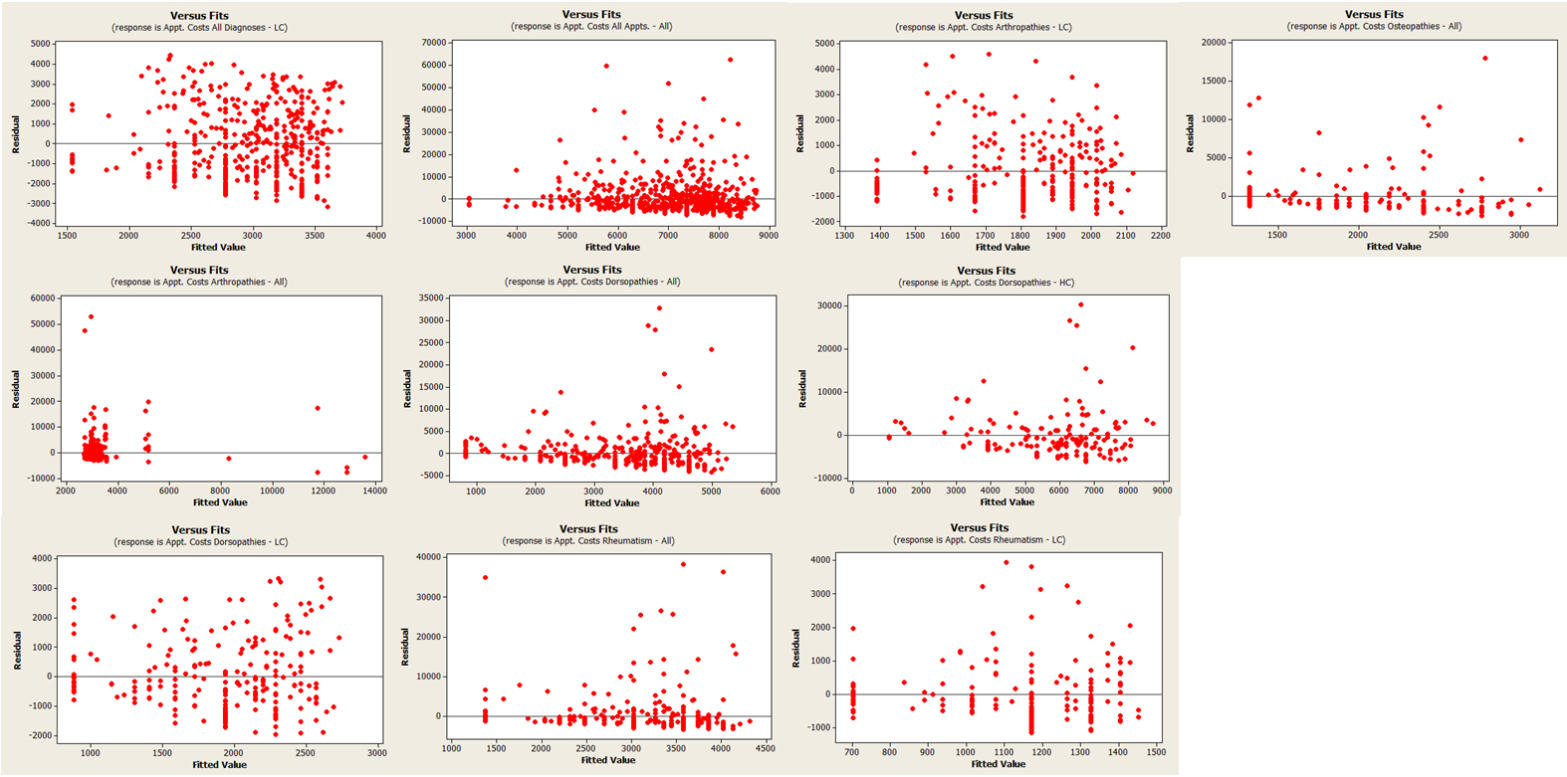

Figure 4: Best Case Scenario Residual versus Fit Plots – Patient Appointment Costs

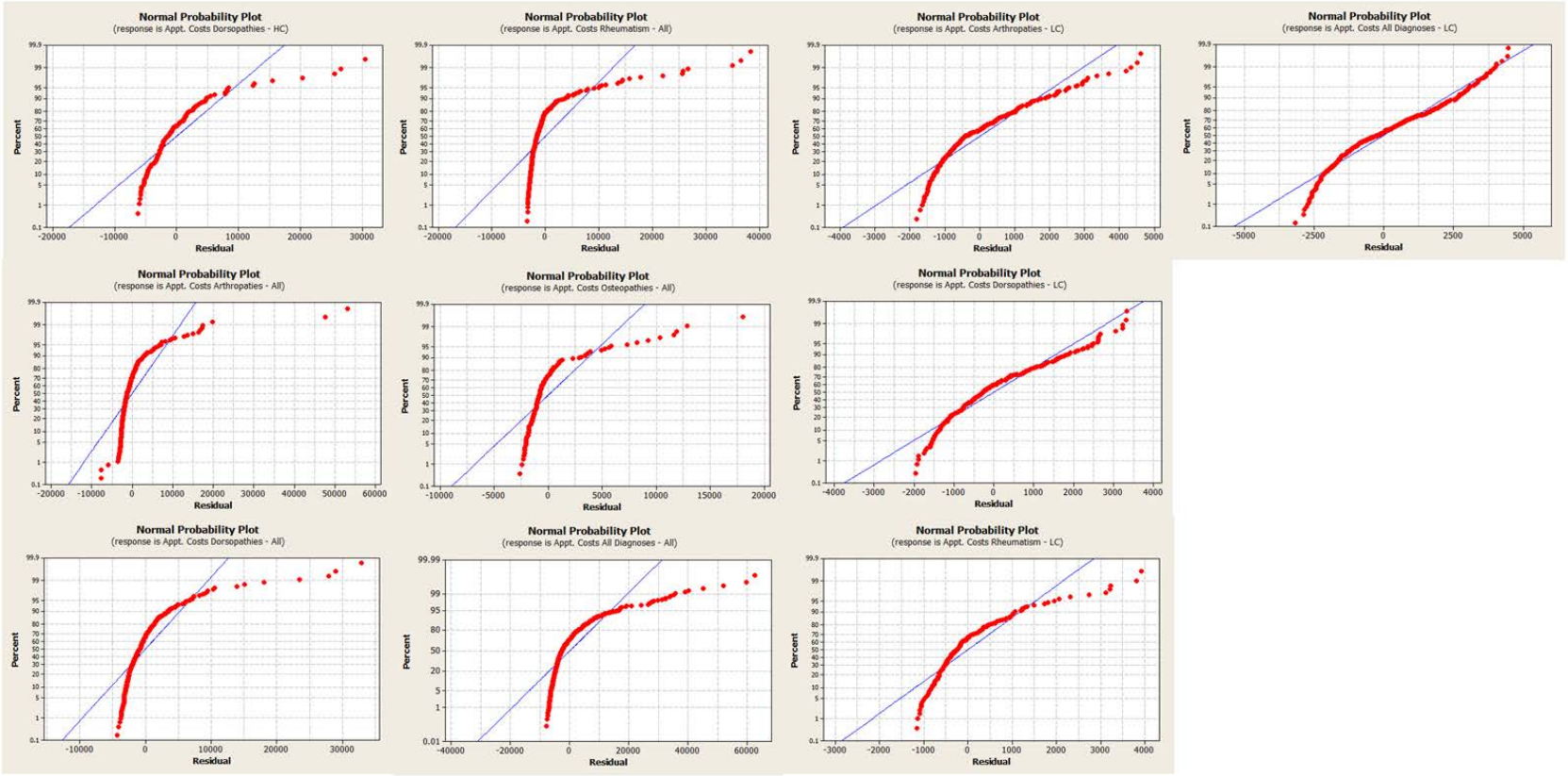

Figure 5: Best Case Scenario Normal Plots for Residuals – Patient Appointment Costs

Worst Case Scenario

Table 8 displays the results for each combination of diagnosis and cost group for the worst case scenario. Cases in which p-values are ≤ 0.05 are highlighted in blue. Similar to the best case scenario, the only case with a p-value ≤ 0.05 for the high cost group is for dorsopathy diagnoses while the low cost group has cases with p-values ≤ 0.05 for all diagnoses with the exception of osteopathies. With both cost groups combined into one group, results reflect p-values are ≤ 0.05 for all cases. Although p-values for each case highlighted in blue ≤ 0.05, R^2 values for each of these equations are very low. Thus, no true conclusions can be drawn about the true impact continuity of care has on patient appointment costs. Figure 6 shows the linear regression graphs of the cases which are statistically significant. Figure 7 and Figure 8 display residual versus fit plots and the normal plots of residuals for the cases in which p-values were less than or equal to 0.05. Similar to that of the best case scenario, in all of the residual versus fit plots, there is a pattern that shows as continuity of care increases, the variability in patient appointment costs also increases. These violate the assumption that there is constant variance along the regression line. Additionally, the normal plots for residuals clearly show that in all cases, the residuals are not normal about the linear regression equation. This violates the second regression assumption of normality. Thus these regression equations are not good models to determine the impact continuity of care has on patient appointment costs in the worst case scenario.

Table 8: Linear Regression Results Worst Case Scenario

Worst Case Scenario	High Cost Group	Low Cost Group	All Patients
Arthropathies	p = 0.656; R^2 = 0.00026 y = -325.65x + 8031	**p = 0.00; R^2 = 0.1565** **y = -2232.5x + 2733.2**	**p = 0.017; R^2 = 0.0159** **y = -909.45x + 2800.5**
Dorsopathies	**p = 0.001; R^2 = 0.066** **y = -7211x + 8391.4**	**p = 0.00; R^2 = 0.1235** **y = -1896.4x + 2780.9**	**p = 0.00; R^2 = 0.0732** **y = -4932.5x + 5485.1**
Rheumatism	p = 0.093; R^2 = 0.0174 y = -4184x + 6733.6	**p = 0.00; R^2 = 0.1058** **y = -1290.8x + 1761.7**	**p = 0.003; R^2 = 0.0262** **y = -3739x + 4666.8**
Osteopathies	p = 0.224; R^2 = 0.0157 y = -1797.5x + 4138.4	p = 0.083; R^2 = 0.035 y = -463.62x + 1168.3	**p = 0.02; R^2 = 0.0296** **y = -1910.5x + 3275.4**
All Diagnoses	p = 0.056; R^2 = 0.0166 y = -11531x + 17883	**p = 0.00; R^2 = 0.1662** **y = -3833.7x + 4330.7**	**p = 0.00; R^2 = 0.1028** **y = -18546x + 14350**

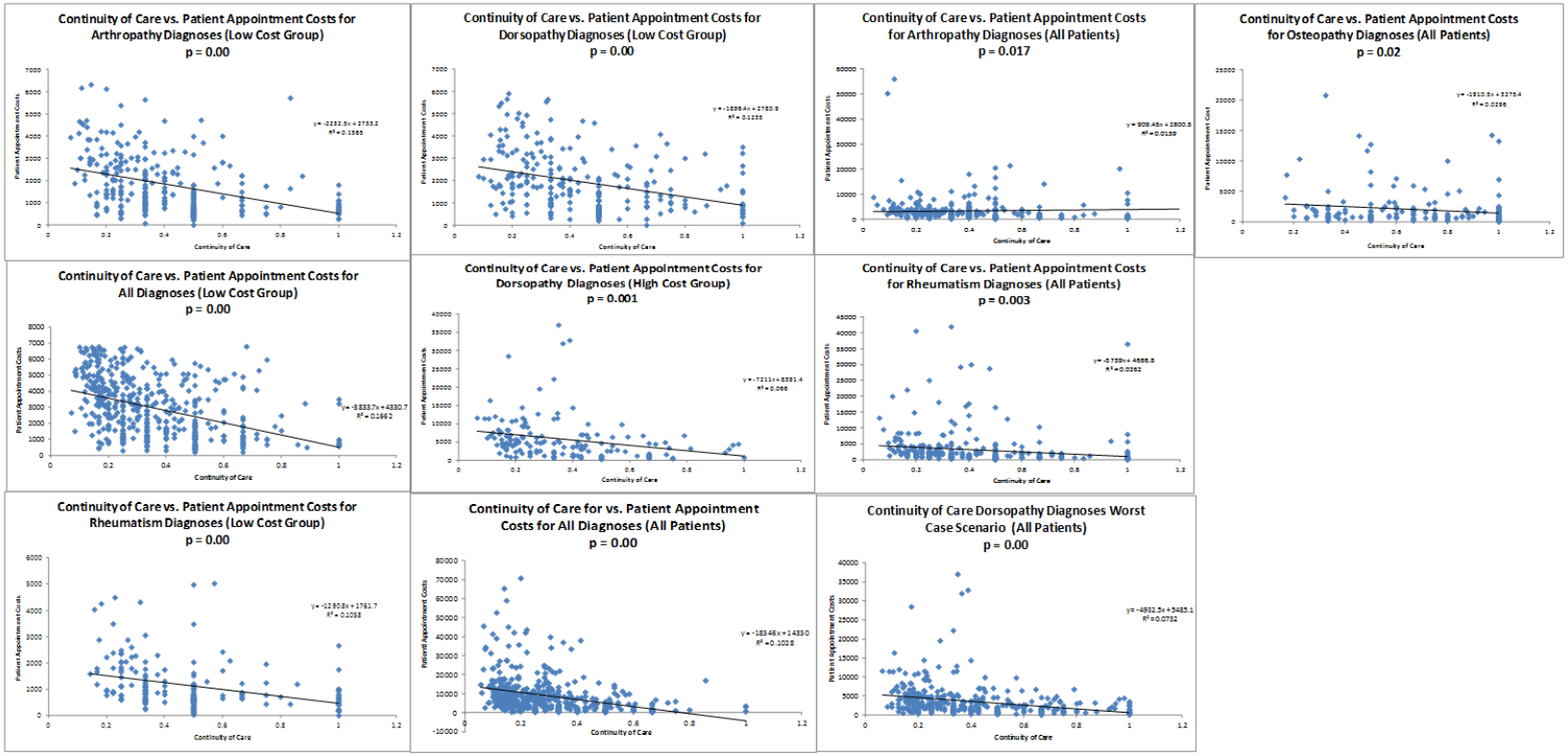

Figure 6: Worst Case Scenario Regression Graphs

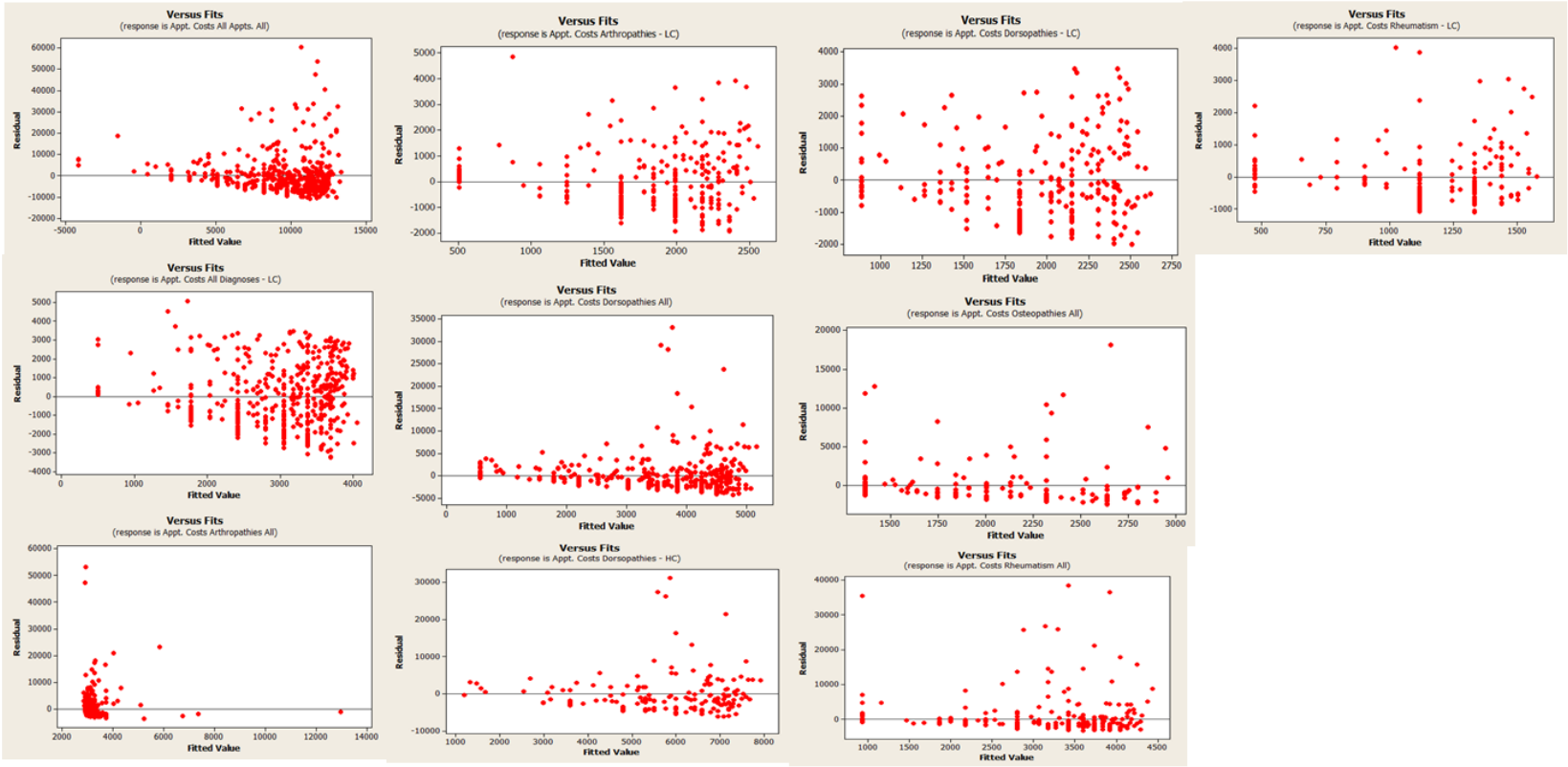

Figure 7: Worst Case Scenario Residual versus Fits Plots – Patient Appointment Costs

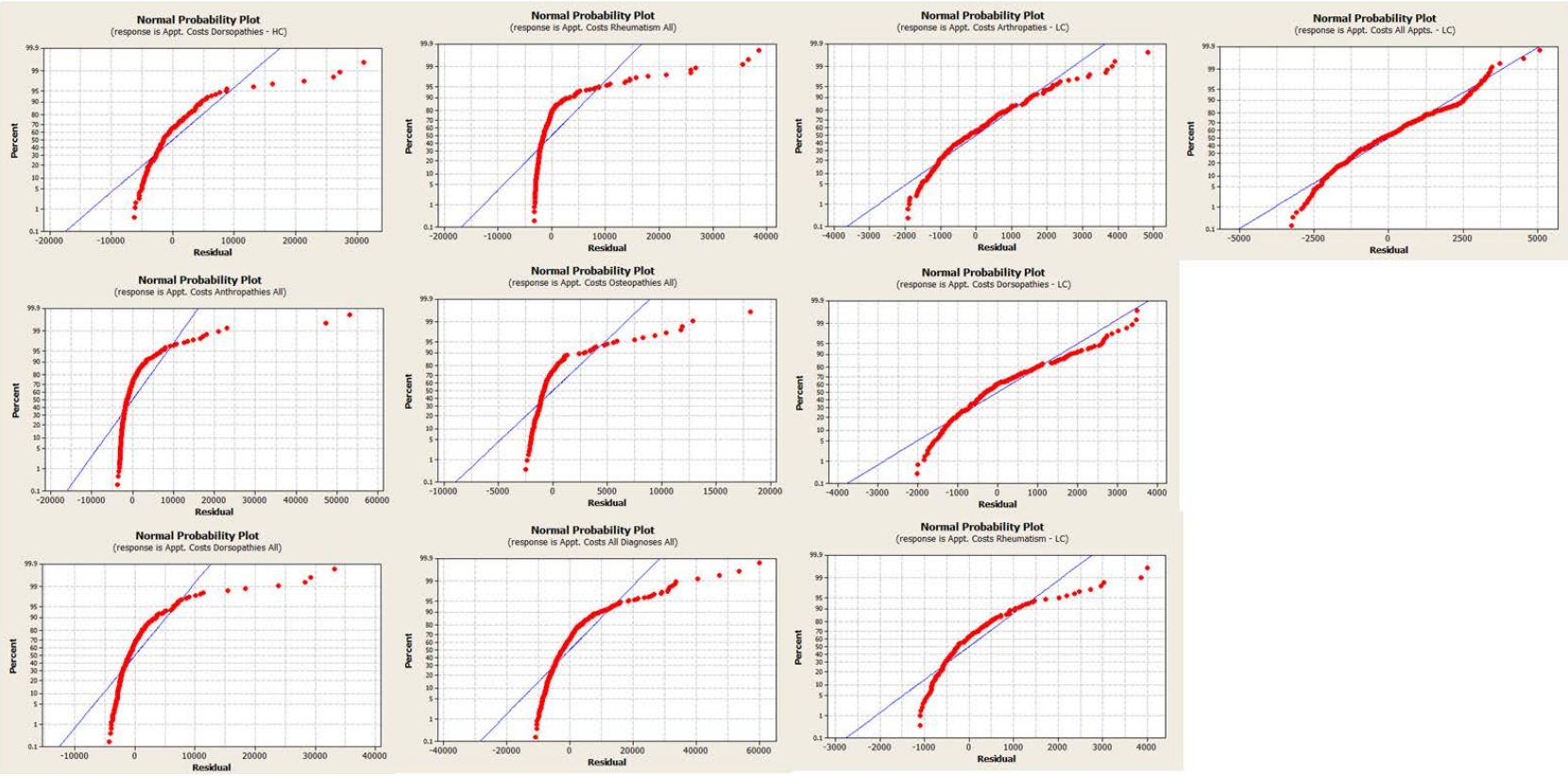

Figure 8: Worst Case Scenario Normal Plot of Residuals – Patient Appointment Costs

Continuity of Care and Patient Availability

Patient availability is defined as the number of days in a calendar year that a patient is on flying status and available to fly. Linear regression is performed on patient availability against continuity of care to determine if continuity of care influences patient availability. This is completed for both the best case and worst case scenarios. The linear regression is performed for each year 2009 through 2012 cumulatively. For example, 2011 will include patient availability calculations using data for 2009, 2010, and 2011. Figures display graphs in which p-values are less than or equal to 0.05.

Best Case Scenario

Table 9 shows the results of the regression analysis that examines continuity of care against patient availability. Cases in which p-values are ≤ 0.05 are highlighted in blue. For the low cost group, 2010 is the only year in which p-value close to 0.05; with p-value of 0.053, its close proximity to the threshold of 0.05 allows it to be highlighted for this study. For the high cost group, cases in which p-values are ≤ 0.05 are in years 2010 and 2011. The graphs in Figure 9 show steeper linear regression lines for the high cost than in the low cost group. Though the p-values are ≤ 0.05, R^2 values are still relatively low, thus no true conclusions can be drawn from the relationship continuity of care has on patient availability. Figure 10 and Figure 11 show the residual versus fit plots and normal plots of residuals for the cases where p-values are less than or equal to 0.05. In the residual versus fits plots, it appears that there is constant variance in all cases; therefore there is no violation of the constant variance assumption of regression. In the normal plots of residuals, the normality assumption appears to be violated for the

low cost group, but not the high cost group. Thus given the normality assumption is violated for the low cost group, this indicates that this regression equation is not a good model to determine the impact continuity of care has on patient availability for the low cost group. Although there are no true violations of the regression assumptions for the high cost group, the low R^2 values still results in no practical significance between continuity of care and patient availability.

Table 9: Patient Availability Regression Results Best Case Scenario

Best Case Scenario	High Cost Group	Low Cost Group
	p = 0.51; R^2 = 0.0031	p = 0.216; R^2 = 0.0077
2009	y = 27.604x + 271.47	y = 44.791x + 255.94
	p = 0.029; R^2 = 0.0304	**p = 0.053; R^2 = 0.0103**
2010	**y = 160.16x + 375.62**	**y = 91.272x + 470.48**
	p = 0.033; R^2 = 0.0201	p = 0.337; R^2 = 0.0021
2011	**y = 229.45x + 533.52**	y = 64.669x + 739.27
	p = 0.227; R^2 = 0.0052	p = 0.36; R^2 = 0.0018
2012	y = 162.43x + 843.95	y = 72.585x + 1059.4

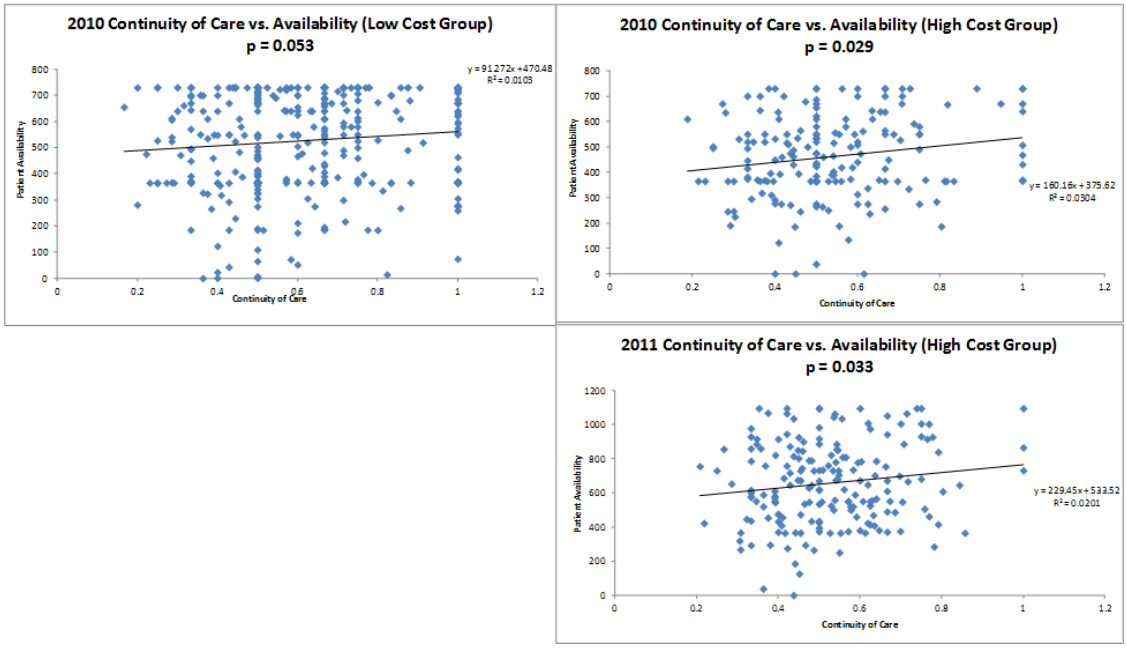

Figure 9: Best Case Scenario Continuity of Care vs. Patient Availability Graph

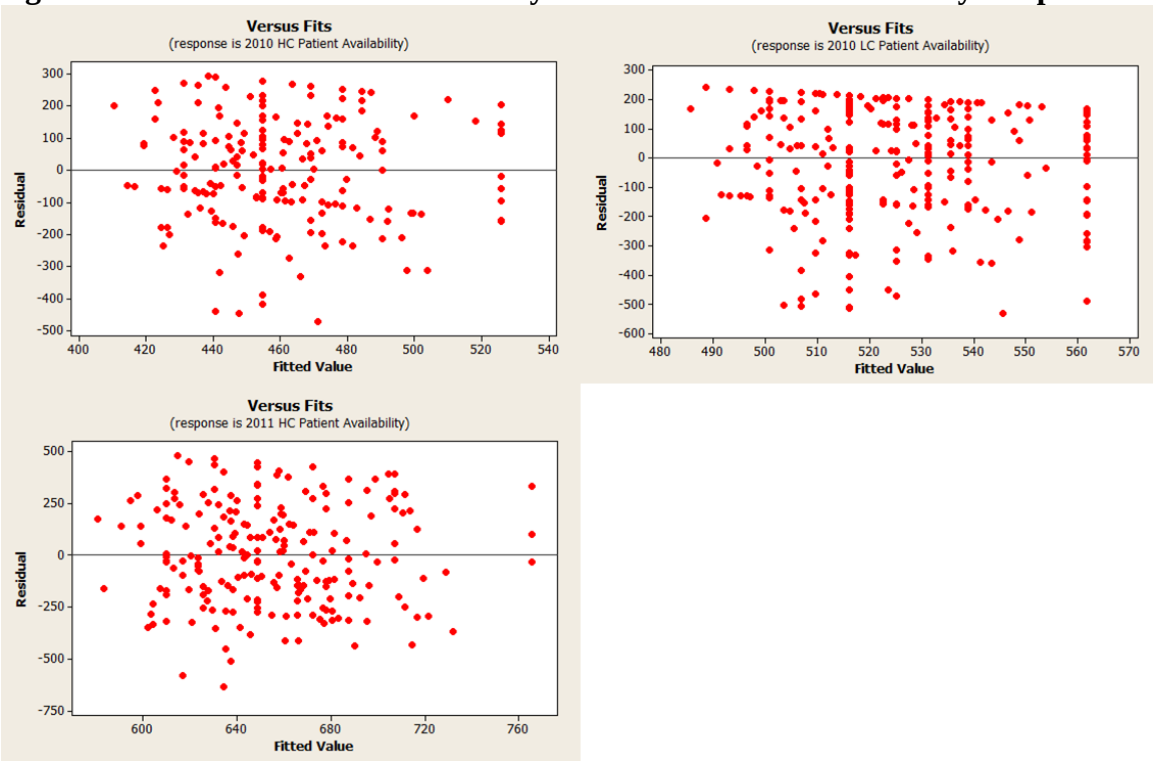

Figure 10: Best Case Scenario Residual Plots - Patient Availability

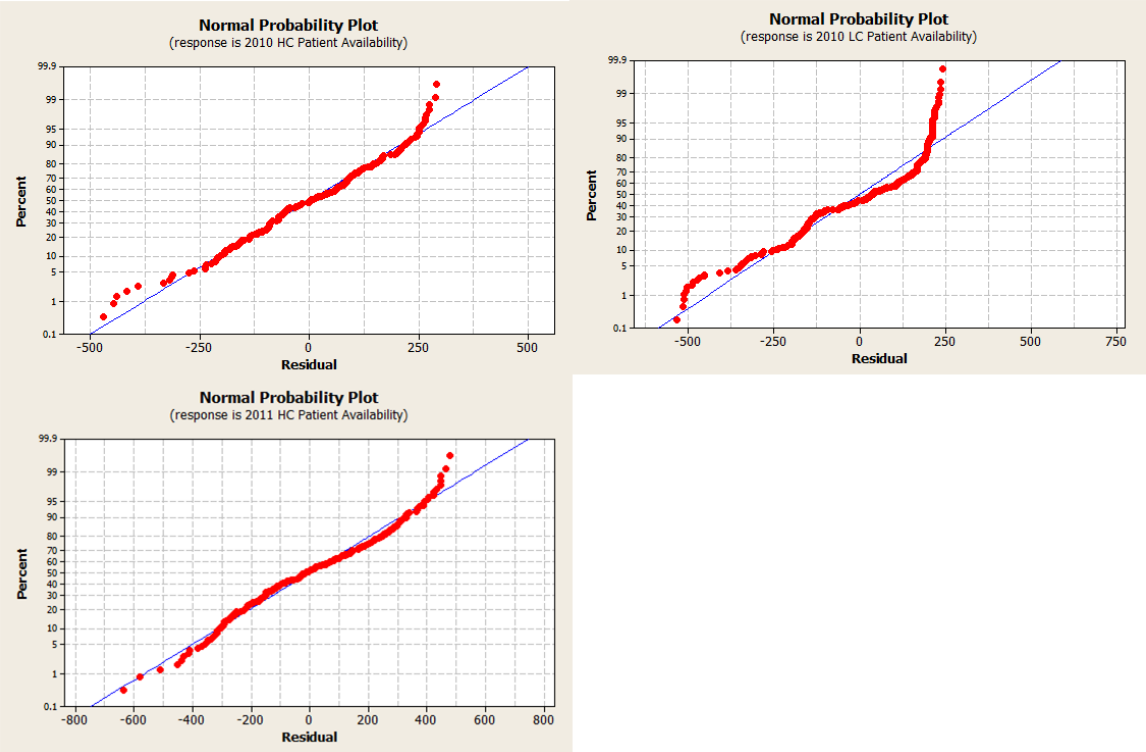

Figure 11: Best Case Scenario Normal Plots of Residuals - Patient Availability

Worst Case Scenario

Table 10 shows the results of the regression analysis that examines continuity of care against patient availability. Cases in which p-values are ≤ 0.05 are highlighted in blue. For the high cost group, the cases with p-values ≤ 0.05 are in years 2010 and 2011. Though the p-values are ≤ 0.05, R^2 values are still relatively low, thus no true conclusions can be drawn from the relationship continuity of care has on patient availability. Figure 12 shows the regression graphs for all cases in which p-values are less than or equal to 0.05. Figure 13 and Figure 14 show the residual versus fits plots and the normal plots of residual for the years in which p-values were less than or equal to 0.05. Similar to that of the best case scenario, the residual versus fits plots for the cases in which p-values are

less than or equal to 0.05 show a constant variance along the regression line. This proves no violation of the regression assumption of constant variance. Also, in the normal plots of residuals, there is no clear violation of normality. Thus these are good models to determine the impact continuity of care has on patient availability in the worst case scenario. However, the low R^2 values still question the strength of the relationship between continuity of care and patient availability.

Table 10: Patient Availability Regression Results Worst Case Scenario

Worst Case Scenario	High Cost Group	Low Cost Group
	p = 0.291; R^2 = 0.0081	p = 0.896; R^2 = 0.00
2009	y = 36.032x + 269.56	y = 4.1548x + 280.66
	p = 0.01; R^2 = 0.0328	p = 0.578; R^2 = 0.00
2010	**y = 138.34x + 403.55**	y = 23.101x + 512.07
	p = 0.004; R^2 = 0.0381	p = 0.924; R^2 = 0.00
2011	**y = 256.05x + 554.29**	y = 5.4858x + 774.3
	p = 0.115; R^2 = 0.00114	p = 0.627; R^2 = 0.00
2012	y = 199.15x + 862.86	y = 32.699x + 1084.7

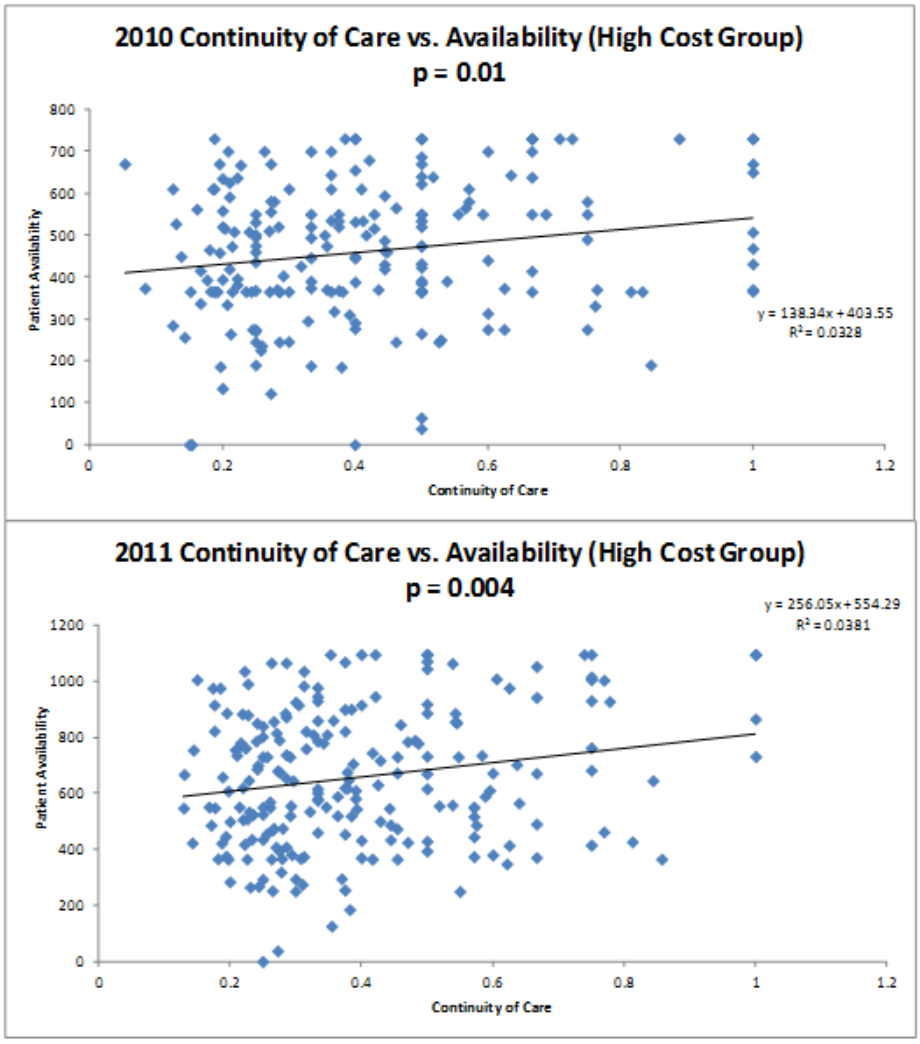

Figure 12: Worst Case Scenario Patient Availability vs. Continuity of Care Graph

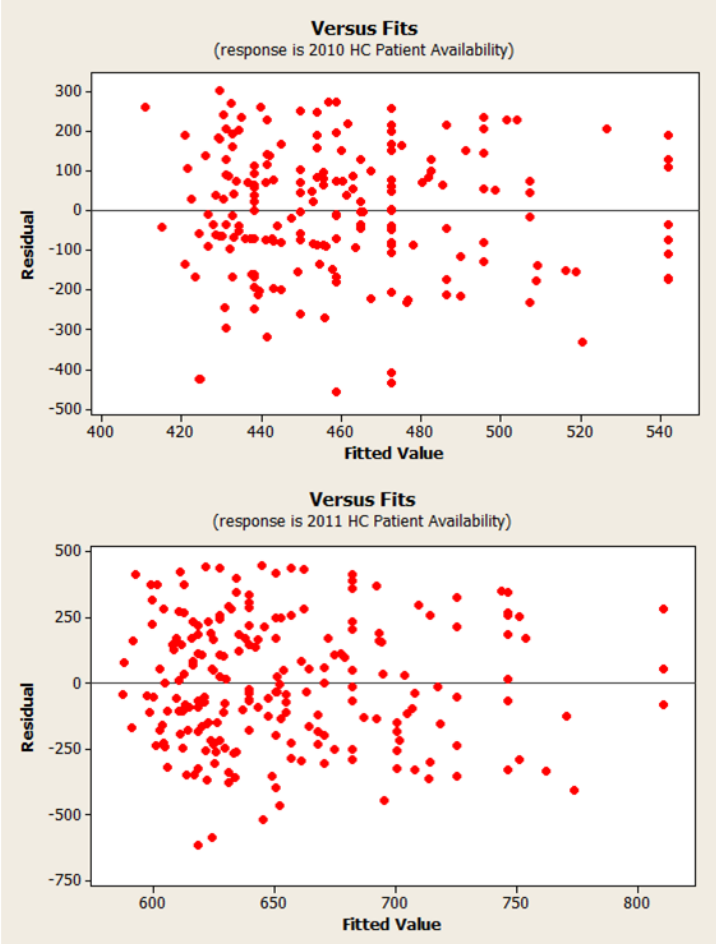

Figure 13: Worst Case Scenario Residual versus Fits Plots - Patient Availability

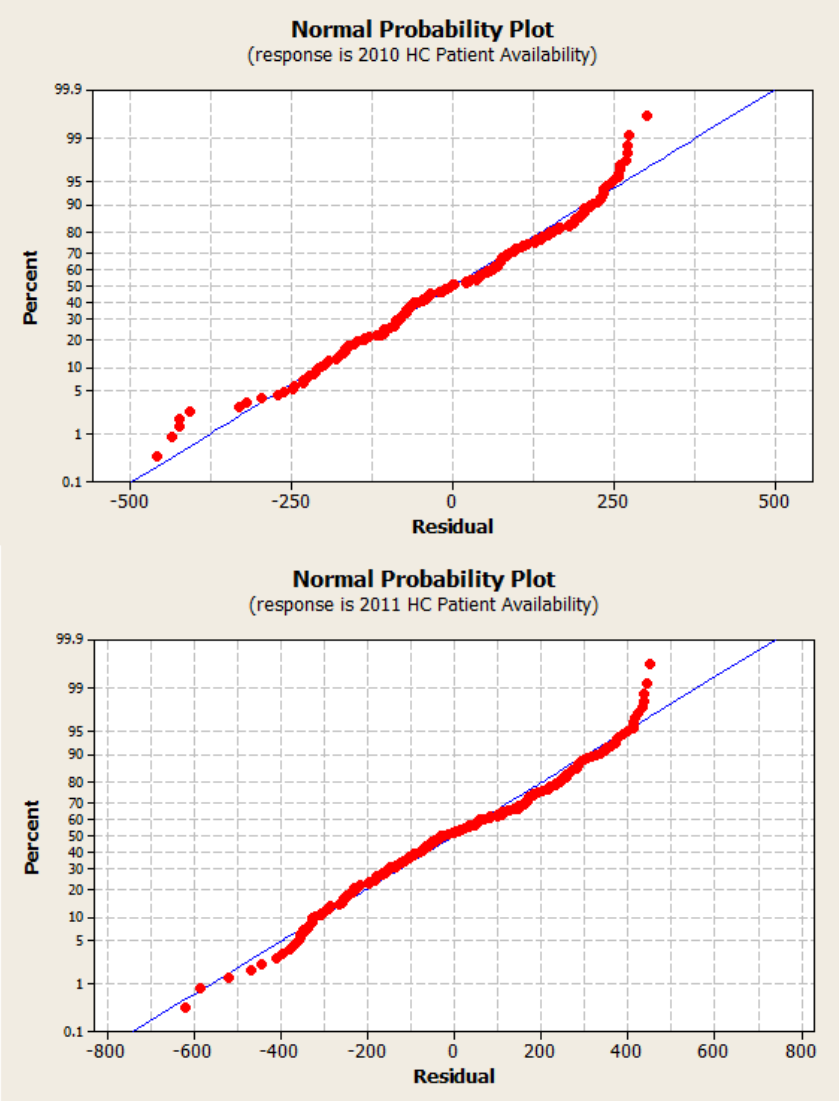

Figure 14: Worst Case Scenario Normal Plots of Residuals - Patient Availability

Conclusion

This chapter analyzes the characteristics that make up different healthcare cost groups within the Air Force flier community. Multivariate and simple linear regression is used to make this determination. The results are unable to conclude that any of the characteristics chosen for this study are predictive of costs. Continuity of care is also analyzed to see how it impacts healthcare cost and patient availability. The analysis

shows continuity of care explains very little of the variability observed in patient appointment costs and patient availability. Further analysis should be performed on a broader population to validate the generalizability of this conclusion.

V. Conclusions and Recommendations

Chapter Overview

Multiple data mining techniques are utilized in this study to determine if continuity of care impacts healthcare costs and patient availability; different cost groups and defining characteristics are also identified in this study. The results provide insights into the current Air Force healthcare system and can be used to improve upon the current model.

Investigative Questions

Investigative Question 1: Are there different cost profiles that make up this population?

This question is answered by calculating the highest cost patients and the percentage of the total cost they contribute. This is done by sorting the population in order by patient total costs and determining the percentage of the subpopulation total costs the highest group accounts for. Using this method, the top 30% of patients are chosen as the high cost group given they account for 70% of the subpopulation total costs. This follows the hypothesized Pareto Rule that a minority percentage of the population is responsible for a majority percentage of healthcare costs. This adds benefit to Air Force healthcare researchers in that the identification of the high cost group enables research to be scoped to target this specific group while still targeting a majority of healthcare costs.

Investigative Question 2: What are the defining characteristics of the different cost populations?

Low Cost Group

Analysis of variance and multivariate linear regression are performed on each characteristic against patient appointment costs in order to determine whether there are characteristics that are predictive of cost. Based on the analysis of variance test, there are no statistically significant costs differences between the different characteristics chosen for this study. Multivariate linear regression show that other/unknown race and the pilot career field have p-values less than 0.05 in each analysis technique; therefore these characteristics are assumed to be predictive of patient appointment cost for the low cost group. Multivariate regression shows that if a patient's race is Other/Unknown, their mean appointment costs are expected to be $824.40 lower. Additionally, if a patient is in the pilot career field, the mean appointment costs are expected to be $450 lower. Adjusted R^2 values of 2.4% indicate these characteristics account for a small portion of the influence of patient appointment costs.

High Cost Group

There are no statistically significant results that show that any personal or organizational factors influence patient appointment costs for the high cost group.

Determining Which Cost Group Patient Belongs To

Binary logistic regression is performed to determine which characteristics predict whether a patient is in the high cost group. Height, weight, fitness test run score, fitness test score, and abdominal circumference provide the best prediction, with odds ratios of 0.85, 1.04, 0.88, 1.15, and 0.84 respectively. The odds ratio is ratio of the probability of an event to the probability of a non event. This is interpreted as for each unit increase, the odds of a patient being in the high cost groups increases by the odds ratio. This

information can be helpful in predicting the likelihood a patient is to end up in the high cost group.

Investigative Question 3: How does continuity of care impact healthcare costs? Does the impact differ for high vs. low cost groups?

Continuity of care is calculated using the percentage of appointments in which a patient meets with the same provider. Simple linear regression is used to determine the relationship between continuity of care and healthcare costs. Due to low R^2 values and violation of regression assumptions, it is concluded that continuity of care explains very little of the variability observed in patient appointment costs.

Investigative Question 4: How does continuity of care impact patient availability?

Simple linear regression is used to determine the impact continuity of care has on patient availability. Due to low R^2 values and violation of regression assumptions, it is concluded that continuity of care explains very little of the variability observed in patient availability.

Significance of Research

The sponsor for this research is the 711th Human Performance Wing (HPW) at Wright Patterson Air Force Base, OH. The Air Force's healthcare model currently consists of primary care managers (PCMs) and PCM teams. Every patient is assigned a PCM and thus subsequently a PCM team. Currently, policy states that it is the goal that each patient meets with their primary care manager (PCM) for 70% of their appointments, or with a member of their PCM team for 90% of their appointments. The research suggests that there are no measureable benefits to cost or patient availability

with increased continuity of care. Knowing this information is beneficial to the Air Force because it can be used to redefine the continuity of care goals and help prioritize other important aspects of healthcare. It is important to note, this study does not investigate the other benefits associated with continuity of care such as improved quality of care, decreased emergency room visits, and decreased number of appointments needed. Before ruling out the need for continuity of care, it may be important to explore these other measures to determine if increased continuity of care adds value to these areas within the Air Force's healthcare system.

Recommendations for Future Research

The next step for furthering research on continuity of care within the Air Force is to expand the research beyond the subpopulation chosen for this study. This research investigates a sub population of Air Force active duty fliers that were off of flying status due to an MSI in July of 2009; expanding this subpopulation to include other non-MSI diagnoses can provide insight into whether the findings presented herein are specific to MSIs only or if the influence continuity of care has on healthcare costs and patient availability are similar for other diagnoses. Furthermore, while none of the personal or organizational characteristics investigated in this study were found to influence patient appointment costs, exploring other characteristics that may be better predictors of costs could provide beneficial insights on drivers of increased healthcare costs.

Summary

This chapter examines each investigative question as stated in the overview chapter, and the conclusions that are drawn based on the results of the analysis. Next the

chapter covers how these conclusions are significant to the sponsor and the recommendations that can be made for further research.

VI. Appendix

Appendix A: IRB Approval Letters

Data Mining Simulation & Optimization of Healthcare Information to Determine Influences on Healthcare Costs & Patient Outcomes
FWR20140109E

1. **Principal Investigator**
   Capt Christina Rusnock, USAF/AFIT, 785-3636x4611, christina.rusnock@afit.edu

2. **Associate Investigators**
   Lt Col Anthony Tvaryanas, 711HPW/HP, 798-3253, anthony.tvaryanas@us.af.mil
   Capt David Wade, USAF/AFIT, 785-3636, david.wade@afit.edu

3. **Facility**
   Secondary data analyses will be conducted at the Air Force Institute of Technology at Wright Patterson Air Force Base, Ohio. Data will be stored on AFIT servers with restricted access to only principal investigator and associate investigators listed in sections 1 and 2.

4. **Objective**
   This research seeks to identify indicators and characteristics of Air Force personnel that contribute to excessive healthcare costs. In order to perform this research, intensive data mining of the health services data will be required. The results of this research can assist in identifying leading indicators of high healthcare costs and processes that contribute to excessive costs. Potential benefits include recommendations on developing manpower requirements per career field based on medical history, managing the staffing levels of healthcare professionals, and other implementing process improvements to the current system in order to reduce long-term Air Force healthcare costs.

5. **Background**
   The cost of the Air Force healthcare system has been growing at a rapid pace [1]. It is hypothesized that the cost is not evenly distributed amongst the entire Air Force population, but instead there exist a large percentage of costs that stems from only a small portion of the population [2].

   Our goal is to implement formal data mining techniques to identify this high-cost sub-population, determine the characteristics of this population, identify predictors of this population, and ultimately optimize healthcare process based on this population. Leveraging the knowledge of the subject matter expert, Lt Col Anthony Tvaryanas, we will be performing data mining techniques, including, but not limited to logistic regression and multivariate linear regression. Based on findings from the data mining, simulation will then be used to capture the variability and emergent system behavior and system dynamics. The information from both the data mining and simulation will be used to build parameters and constraints to formulate an optimization problem. Our secondary goal is to conduct a process improvement study on a military clinic located at Wright Patterson AFB implementing formal discrete-event simulation

Data Mining Simulation & Optimization of Healthcare Information to Determine Influences on Healthcare Costs & Patient Outcomes
FWR20140109E
AFRL IRB Approval Valid from 4 September 2014

- Is there a small portion of the AF population that drives healthcare costs or patient availability? (in a given year, cumulative?)
- What was the lost duty time (for special duty personnel) associated with disease conditions?
- What disease conditions are correlated within individuals?
- How are disease conditions and health services utilization correlated within individuals and their associated beneficiaries?
- Are there unique characteristics that make up this sub population (AFSC, deployment status, fitness, geo location, age)?
- What characteristics of patient care (continuity, provider type) result in lower cost or increased availability for this sub population?
- How does the cost benefit of continuity of care change in chronically ill patients versus patients with acute illnesses?
- Does continuity of care impact healthcare costs? Does the impact differ for high cost vs. low cost populations?
- Does continuity of care impact patient availability? Does the impact differ for high vs. low cost populations?
- Does high cost diagnostic test increase availability? Does higher cost treatments increase availability?

The research project will also utilize centralized medical databases maintained by AF/SG6 to conduct a simulation study on the clinics located at Wright Patterson AFB. Ms. Genny Maupin will be the 3rd party data broker with AF/SG6 and will be responsible for de-identifying the data prior to transferring it to the PIs. The objective of this simulation is to identify the baseline process of the current system, identify where the bottlenecks occur, and identify the relative cost-effectiveness of current and alternative personnel staffing levels and processes as well as mitigate the patient wait time. Data will be obtained from the AFPC Personnel database, Defense Enrollment Eligibility Reporting System (DEERS), the Aviation Safety Information Management System (ASIMS), the Air Force Fitness Management System (AFFMS), the Cardiac Risk Assessment and Management (CRAM) database, the Aeromedical Information Management Waiver Tracking System (AIMWTS) and the Military Health System (MHS) data Mart (M2), which contains on-base outpatient clinic visits for the entire Air Force. Data collected will include demographics (i.e. age and gender), encounter dates, duty not including flying information (DNIF) codes, fitness restrictions, physical fitness test scores, cardiac risk scores, waivers for flyers, duty location and Air Force Specialty Code (AFSC) and procedure and diagnosis codes to assess the utilization of services and investigate the associated disease conditions. Data collected for the simulation study will include the patient category (i.e. military, civilian and dependent), appointment type, diagnosis codes, appointment status, clinics located at Wright Patterson AFB and provider type to access the probabilities of a particular type of patient visiting the clinic.

Data Mining Simulation & Optimization of Healthcare Information to Determine Influences on Healthcare Costs & Patient Outcomes
FWR20140109E
AFRL IRB Approval Valid from 4 September 2014

techniques to identify opportunities the military healthcare system can implement to advance the efficiency and effectiveness of our military healthcare system. The discrete-event simulation will map the process of a base level health clinic and identify the relative cost-effectiveness of current and alternative personnel staffing levels and processes. The results of our finding will aid the clinic to minimize staffing cost, identify bottlenecks in the system, minimize patient wait time, maximize the utilization of its medical personnel, and ultimately increase the efficiency and effectiveness the military healthcare system.

6. **Impact to Air Force Mission**

Given that the DoD health care costs are growing more than twice as fast as economy-wide medical inflation, there is reason for serious concern that increasing health care expenditures will reduce resource availability for other important defense programs and undermine the overall capability of the U.S. military. Consequently, there is an urgent need to bring health care costs into a sustainable range, and all aspects of the defense health portfolio must be subject to critical review—and aerospace medicine can be no exception.

7. **Experimental Plan**

a. Equipment:
Existing computers that have Arena, MiniTab, JMP, and Microsoft Office 2007 software installed will be used. These computers require CAC enabled access.

b. Subjects:
Active duty Air Force special duty personnel and their beneficiaries. Additional analyses focused on Wright Patterson AFB population utilizing the clinics at Wright Patterson AFB.

c. Duration:
Timeframe for data analysis is June 2014 to June 2016; timeframe for competition of the study is approximately 24 months.

d. Description of experiment, data collection, and analysis:
The research project will utilize centralized medical databases maintained by AF/SG6 to perform a cross-sectional audit over a 10-year period (CY03-CY13) at overall Air Force healthcare system. As such, all appropriate data use agreements will be obtained prior to acquiring data. Data will be pulled by Ms. Genny Maupin; personal identifiers will be stripped before data is forwarded for analysis will be performed. The objective of this audit is to find the characteristics that define the population of active duty Air Force members and their dependents that make up the highest percentage of Air Force healthcare costs. In addition, this audit will look to find the differences in recovery time for persons with musculoskeletal injury that had the highest continuity with the health care provider. Specific information will be elicited on the following questions:

The study will also collect time data (i.e. the time it takes for a process to be completed) at the clinic to develop probability distributions for the simulation.

Regression analysis will be performed to determine which illnesses or characteristics of patients make up the highest percentage of healthcare cost annually, and how these have changed over the years. To tailor our research, it will be necessary to have available as much information as possible to determine which characteristics correlate with costs.

Since PII and PHI will be collected from over a ten year period and involve up to 100,000 participants, it is unrealistic to obtain informed consent for the PII and HIPAA authorization for the PHI. Hence, both a waiver of informed consent and a waiver of HIPAA authorization are needed.

Accordingly, a waiver of informed consent is here forth requested. Per 32 CFR 219.116(d), "an IRB may approve a consent procedure which does not include, or which alters, some or all of the elements of informed consent set forth in this section, or waive the requirements to obtain informed consent provided the IRB finds and documents that:

i. The research involves no more than minimal risk to the subjects;
ii. The waiver or alteration will not adversely affect the rights and welfare of the subjects;
iii. The research could not practicably be carried out without the waiver or alteration; and
iv. Whenever appropriate, the subjects will be provided with additional pertinent information after participation."

e. In the case of this protocol, the research involves no more than minimal risk to the subjects as indentifying information will only be used to link data from disparate information sources and will then be removed from the dataset. Additionally, the research could not be practically carried out as informed consent from each research subject is not possible due to the difficulty in locating each subject given the time frame of interest (i.e., 10-year period), the size of the sample (which will include several thousand subjects), and the short time frame over which the study will be conducted.

f. Safety monitoring:
Not applicable, as study is minimal risk.

g. Confidentiality protection:
Computers used for data management will be located at the Air Force Institute of Technology at Wright Patterson AFB, OH in building 640. The PI and AI's computers require appropriate access, i.e. Common Access Card (CAC). In addition, all personal identifiers will be stripped prior to analysis, once data are merged, and random numbers will be assigned to each individual in the dataset. No key or code will be kept linking the random numbers to the Personally Identifiable Information (PII) or Protected Health Information (PHI). Data will not be analyzed or investigated until the identifiers have

been stripped. The associate investigators have all completed CITI training, and possess active Secret or Top Secret security clearances. Data will be deleted once the study is complete (approximately 24 months).

9. Risk Analysis

The main risk to subjects is potential release of PII and PHI. This risk will be mitigated by the procedures in 8f. Another risk to subjects is the release of findings that could potentially shed a negative light on the career fields studied. All reports and presentations will be routed through the appropriate Public Affairs (PA) and Scientific and Technical Information (STINFO) channels prior to release outside the organization.

10. References

1. Harrison, Todd. "The New GuNs Versus BuTTer DeBaTe." Center for Strategic and Budgetary Assesments (2010).
2. Weinberg, Myrl. "In health-care reform, the 20-80 solution." The Providence Journal (2009).

11. Attachments

a. Capt Christina Rusnock, CV
b. Capt Christina Rusnock, CITI Training Certificate
c. Lt Col Anthony Tvaryanas, CV
d. Lt Col Anthony Tvaryanas, CITI Training Certificate
e. Capt David Wade, CV
f. Capt David Wade, CITI Training Certificate

Data Mining Simulation & Optimization of Healthcare Information to Determine Influences on Healthcare Costs & Patient Outcomes
FWR20140109E
AFRL IRB Approval Valid from 4 September 2014

DEPARTMENT OF THE AIR FORCE
AIR FORCE RESEARCH LABORATORY
WRIGHT-PATTERSON AIR FORCE BASE OHIO 45433

MEMORANDUM FOR USAF/AFIT (CAPT CHRISTINA RUSNOCK)

FROM: 711 HPW/IR (AFRL IRB)

SUBJECT: IRB approval for the use of human volunteers in research

1. Protocol title: Data Mining Simulation & Optimization of Healthcare Information to Determine Influences on Healthcare Costs & Patient Outcomes
2. Protocol number: FWR20140109E
3. Protocol version: 1.01
4. Risk: N/A
5. Approval date: 4 September 2014
6. Expiration date: N/A
7. Scheduled renewal date: N/A
8. Type of review: Exempt
9. Assurance Number and Expiration Date: N/A
10. CITI Training: Completed
11. The above protocol has been reviewed and determined to be exempt from IRB oversight. The objective of the study is to identify indicators and characteristics of Air Force personnel that contribute to excessive healthcare costs, in order to have data upon which to provide recommendations about manpower requirements, staffing levels and other healthcare system improvements that will ultimately reduce the cost of providing healthcare. Up to 100,000 subject health records are expected to be included in the retrospective health care record data mining effort. Access to PHI data bases will be provide by SG6 and proper data use agreements will be in place. The data will be collected by a disinterested third party (Ms. Gen Maupin) who will provide a fully de-identified database to the Principle Investigator for analysis. No identifiable data will be accessed by or recorded by the researchers. **Amendments:** Changes to Section 7 part d which include a change to the focus area of continuity of healthcare providers with persons with musculoskeletal injury, three additional research questions on impact of continuity of care on cost and patient availability, four new databases that data will be collected from, six new data fields that will be obtained and analyzed. This protocol therefore meets the criteria for exemption in accordance with 32 CFR 219.101 (b)(4) which exempts "Research, involving the collection or study of existing data, documents,

records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects."

12. HIPAA authorization is required to access PHI from AHLTA for these research purposes. HIPAA waiver is granted having found this to be minimal risk study, wherein the study could not be conducted without access to the PHI, consent could not practicably be obtained, PHI accessed is limited to the minimal number of records as is needed to meet the research goal, and adequate privacy/security safeguards are in place.

13. FDA regulations do not apply since no drugs, supplements, or unapproved medical devices will be used in this research.

14. This exemption applies only to the requirements of 32 CFR 219, DoDI 3216.02, AFI 40-402, and related human research subject regulations.

15. With this approval comes the expectation that the Principle Investigator has the funding to fully execute the protocol. Partial protocol funding, particularly with Greater than Minimal Risk studies, should prompt a re-examination of the protocol by both the Principle Investigator and the IRB with specific emphasis on the risk-benefit evaluation.

16. Any serious adverse event or issues resulting from this study should be reported immediately to the IRB. Amendments to protocols and/or revisions to informed consent documents must have IRB approval prior to implementation. Please retain both hard copy and electronic copy of the final approved protocol and informed consent document.

17. The IRB must be notified if there is any change to the design or procedures of the research to be conducted. Otherwise, no further action is required. All inquiries and correspondence concerning this protocol should include the protocol number and name of the primary investigator.

18. For questions or concerns, please contact the IRB administrator, Lt Eric Fergueson at william.fergueson@us.af.mil or (937) 904-8094. All inquiries and correspondence concerning this protocol should include the protocol number and name of the primary investigator.

LONDON.KIM.ELIZABETH.1155556370

Digitally signed by LONDON.KIM.ELIZABETH.1155556370
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=LONDON.KIM.ELIZABETH.1155556370
Date: 2014.09.04 11:28:19 -04'00'

KIM E. LONDON, JD, MPH, CIP
Chair, AFRL IRB

1st Indorsement, USAF/AFIT (CAPT CHRISTINA RUSNOCK), Memo, 4 September 2014, IRB Approval for the Use of Humans in Research, Expedited Review, Exempt Approval, Protocol Number FWR20140109E

MEMORANDUM FOR 711 HPW/IR (KIM LONDON)

I have reviewed the hardcopy and electronic records and found them to be complete and accurate.

FERGUESON.WILLIAM.ERIC.1296513071

Digitally signed by FERGUESON.WILLIAM.ERIC.1296513071
DN: c=US, o=U.S. Government, ou=DoD, ou=PKI, ou=USAF, cn=FERGUESON.WILLIAM.ERIC.1296513071
Date: 2014.09.04 10:57:52 -04'00'

W. ERIC FERGUESON, 2LT, USAF
Lead Administrator, AFRL IRB

Appendix B: Best Case Scenario for Continuity of Care vs. Patient Appointment Costs Graphs for p-values > 0.05

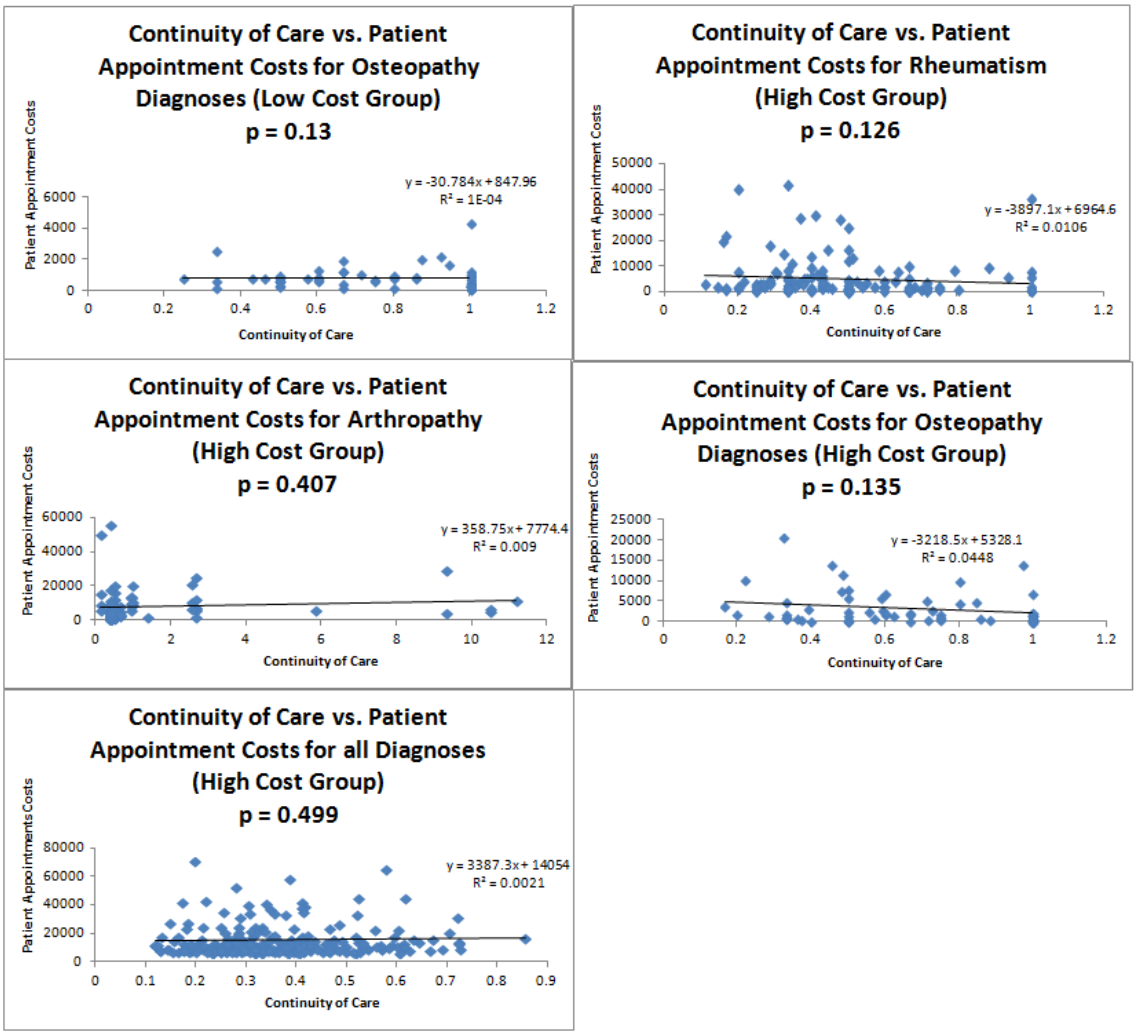

Figure 15: Best Case Scenario Regression Graphs (Cases p > 0.05)

Appendix C: Worst Case Scenario for Continuity of Care vs. Patient Appointment Costs Graphs for p-values > 0.05

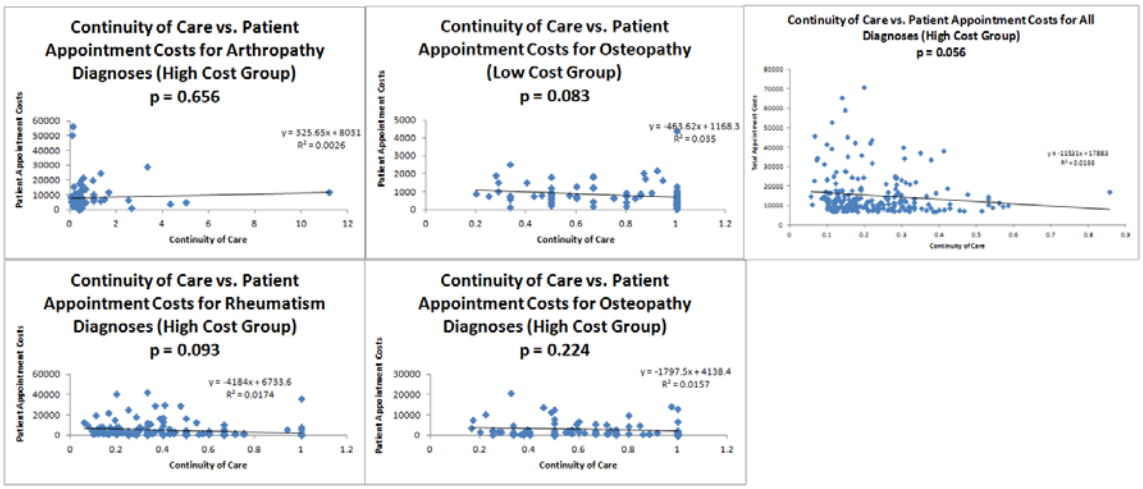

Figure 16: Worst Case Scenario Regression Graphs (Cases p > 0.05)

Appendix D: Best Case Scenario Continuity of Care vs. Patient Availability Graphs for p-values > 0.05

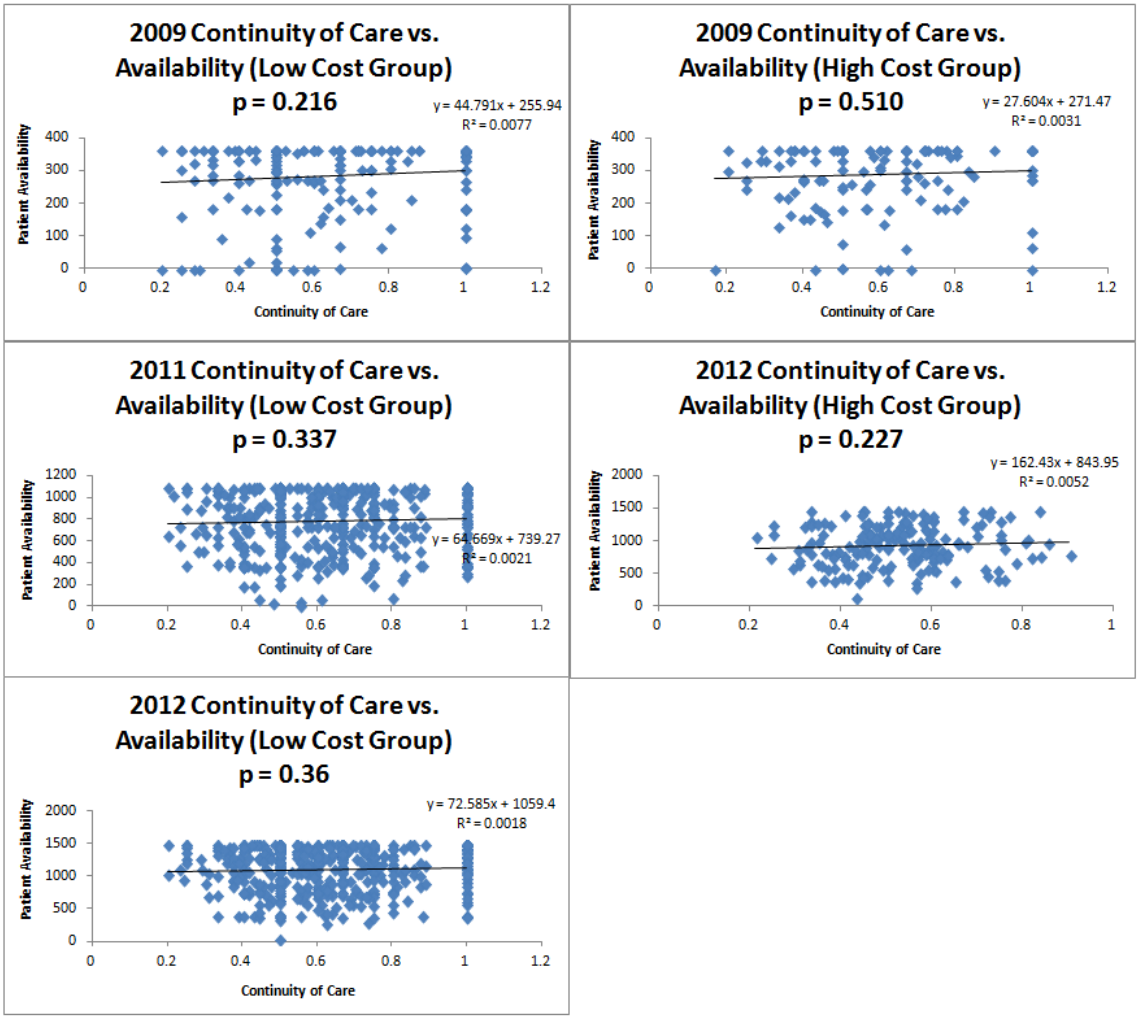

Figure 17: Best Case Scenario Patient Availability vs. Continuity of Care Graph (Cases p > 0.05)

Appendix E: Worst Case Scenario Continuity of Care vs. Patient Availability Graphs for p-values > 0.05

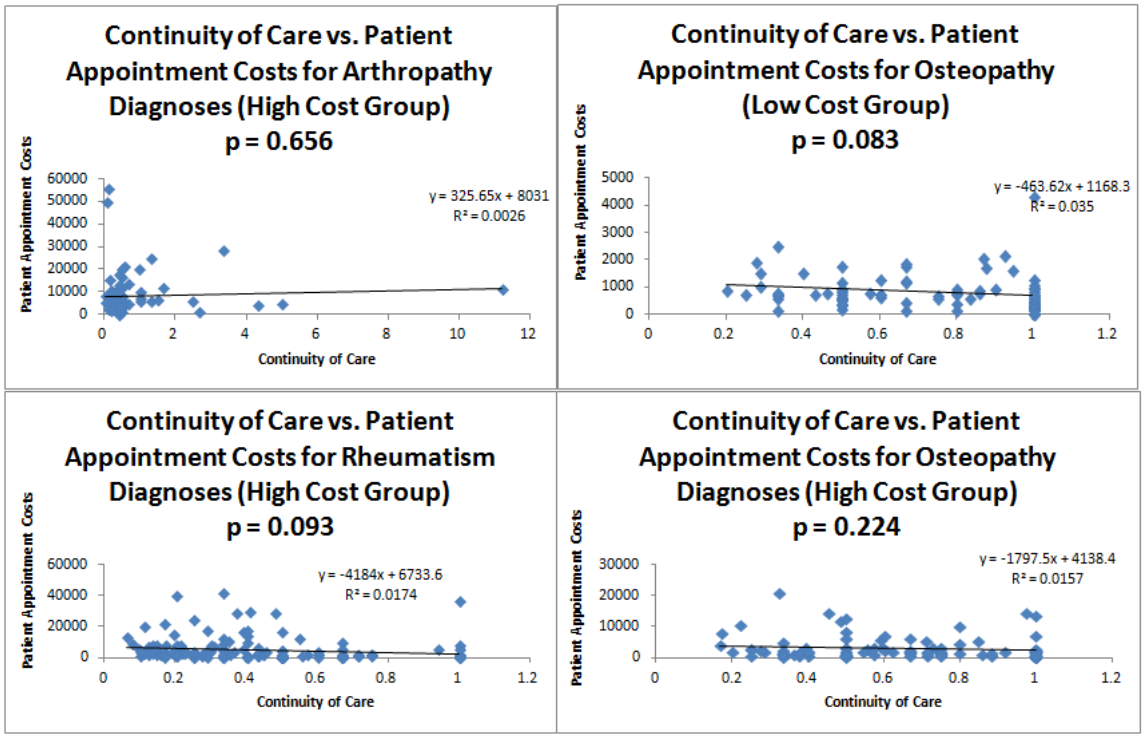

Figure 18: Worst Case Scenario Patient Availability vs. Continuity of Care Graph (Cases p > 0.05)

Bibliography

Anderson, L. H., Flottemesch, T. J., Fontaine, P., Solberg, L. I., & Asche, S. E. (2012). Patient Medical Group Continuity and Healthcare Utilization. *The American Journal of Managed Care, 18(8)*, 450-457.

Balasubramanian, H., Banerjee, R., Denton, B., Naessens, J., & Stahl, J. (2010). Improving Clinical Access and Continuity through Physician Panel Redesign. *Journal of General Internal Medicine, 25(10)*, 1109-1115.

Balasubramanian, H., Banerjee, R., Gregg, M., & Denton, B. T. (2007 Winter). Improving Primary Care Access Using Simulation Optimization. *IEEE*, (pp. 1494-1500).

Bjorkelund, M. C., Maun, A., Murante, A. M., Hoffmann, K., De Maeseneer, J., & Farkas-Pall, Z. (2013). Impact of Continuity on Quality of Primary Care: From the Perspective of Citizens' Preferences and Multimorbidity-Position Paper of the European Forum for Primary Care. *Quality in Primary Care 21(3)*, 193-204.

Bose, I., & Mahapatra, R. K. (2001). Business Data Mining--A Machine Learning Perspective. *Information & Management, 39.3*, 211-225.

Brannigan, M. (1999). Quintiles Seeks Mother Lode in Health "Data Mining". *Wall Street Journal, March 2,1*.

De Maeseneer, J. M., De Prins, L., Gosset, C., & Heyerick, J. (2003). Provider Continuity in Family Medicine: Does it Make a Difference for Total Health Care Costs? *The Annals of Family Medicine, 1(3)*, 144-148.

Dolfini-Reed, M., & Jebo, J. (2000). *The Evolution of the Military Health Care System: Changes in Public Law and DoD Regulations.* Alexandria, VA: Center for Naval Analyses.

Farrell, D. (2008). *Accounting for the cost of US health care: A new look at why Americans spend more.* McKinsey & Company.

Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *Al Magazine, 17(3)* , 37.

Gilmer, T. P., O'Connor, P. J., Rush, W. A., Crain, A. L., Whitebird, R. R., Hanson, A. M., & Solber, L. I. (2005, 28(1)). Predictors of Health Care Costs in Adults with Diabetes. *Diabetes Care*, 59-64.

Gilmer, T. P., O'Connor, P. J., Rush, W. A., Crain, A. L., Whitebird, R. R., Hanson, A. M., & Solberg, L. I. (2005). Predictors of Health Care Costs in Adult with Diabetes. *Diabetes Care,* 59-64.

Green, L. V., Savin, S., & Murray, M. (2007). Providing Timely Access to Care: What is the Right Patient Panel Size? *Joint Commission Journal on Quality and Patient Safety, 33(4),* 211-218.

Hamrock, E., Paige, K., Parks, J., Scheulen, J., & Levin, S. (2012). Discrete Event Simulation for Healthcare Organizations: A Tool for Decision Making. *Journal of Healthcare Management/American College of Healthcare Executives, 58(2),* 110-24.

Harrell, F. E. (2001). *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis.* Springer.

Harrison, T. (2010). *The New Guns Versus Butter Debate.* Washington: Center for Strategic and Budgetary Assesments.

Jacobson, S. H., Hall, S. N., & Swsher, J. R. (2006). Discrete-Event Simulation of Health Care Systems. *Patient Flow: Reducing Delay in Healthcare Delivery,* 211-252.

Kincade, K. (1998). Data Mining: Digging for Healthcare Gold. *Insurance & Technology, 23(2),* IM2-IM7.

Koh, H. C., & Tan, G. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management Vol, 19(2),* 65.

Kristjansson, E., Hogg, W., Dahrouge, S., Tuna, M., Mayo-Bruinsma, L., & Gebremichael, G. (2013). Predictors of Relational Continuity in Primary care: Patient, Provider and practice factors. *BMC Family Practice,* 72.

Kurth, W. T., Glynn, A. M., Gaziano, K. A., Berger, K., & Robins, J. M. (2006). Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting Under Conditions of Nonuniform Effect. *American Journal of Epidemiology, 163(3),* 262-270.

Lv, X., Wu, Z., Jiang, C., Li, Y., Yang, X., Zhang, Y., & Zhang, N. (2011). Complication Risk of Endovascular Embolization for Cerebral Arteriovenous Malformation. *European Journal of Radiology 80(3),* 776-779.

Macinko, J., Starfield, B., & Shi, L. (2007). Quantifying the Health Benefits of Primary Care Physician Supply in the United States. *International Journal of Health Services, 37(1),* 111-126.

Mainous, A. G., & Gill, J. M. (1998). The Importance of Continuity of Care in the Likelihood of Future Hospitalization: Is Site of Care Equivalent to a Primary Clinician? *American Journal of Public Health, 88(10),* 1539-1541.

Mitchell, R. (2013, March 26). Slow Progress on Efforts to Pay Docs, Hospitals For 'Value,' Not Volume. *Kaiser health news*.

PCMH. (2014). Retrieved May 23, 2014, from National Committe for Quality Assurance (NCQA): http://www.ncqa.org/Programs/Recognition/RelevanttoAllRecognition/RecognitionTraining/PCMH2014Standards.aspx

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., & Hammond, W. E. (1997). Medical Data Mining: Knowledge Discover in a Clinical Data Warehouse. *Proceedings of the AMIA Annual Fall Symposium, American Medical Informatics Association,* 101.

Ramachandran, S., Erraguntla, M., Mayer, R., & Benjamin, P. (2007). Data Mining in Military Health Systems-Clinical and Administrative Applications. *Automation Science and Engineering* (pp. 158-163). IEEE International Conference.

Reid, R. J., Fishman, P. A., Onchee, Y., Ross, T. R., Tufano, J. T., Soman, M. P., & Larson, E. B. (2009). Patient-Centered Medical Home Demonstration: A Prospective, Quasi-Experimental, Before and After Evaluation. *The American Journal of Managed Care, 15(9),* 71-87.

Schieber, S. J., Bilyeu, C. D., Hardy, D. R., Katz, M. R., Kennelly, B. B., & Warshawsky, M. J. (2009). The Unsustainable Cost of Health Care. *Social Security Advisory Board, Washington DC,* 1-43.

Schuerenberg, B. K. (2003). An Information Excavation. *Health Data Management, Vol. 11(6),* 80-82.

Sharma, A., & Mansotra, V. (2014). Emerging Applications of Data Mining for Healthcare Management-A Critical Review. *Computing for Sustainable Global Development (INDIACom).* International Conference on. IEEE.

Srinivas, K. B., Kavihta, R., & Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. *International Journal on Computer Science and Engineering (IJCSE) 2, no. 02*, 250-255.

Tan, G., & Koh, H. C. (2011). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management Vol 19.2*, 65.

Tantau, C. (2009). Accessing Patient-Centered Care Using the Advanced Access Model. *The Journal of Ambulatory Care Management 32, no. 1*, 32-43.

Tvaryanas, A. (2014). MD, Aerospace Medicine and Technical Advisor, 711th Huamn Systems Integration Directorate. (D. Wade, Interviewer)

Weinberg, M. (2009). In Health Care Reform, the 20-80 Solution. *Providence Journal*.

Weiss, L. J., & Blustein, J. (1996). Faithful Patients: The Effect of Long-Term Physician-Patient Relationships on the Costs and Use of Health Care by Older Americans. *American Journal of Publich Health 86, no. 12*, 1742-1747.

Wong, W.-K. (2004). *Data Mining for Early Disease Outbreak Detection.*

Yamane, G. K. (2006). Cancer Incidence in the US Air Force: 1989-2002. *Aviation, Space and Environmental Medicine 77.8*, 789-794.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
26-12-2014	Master's Thesis	October 2013 – December 2014

TITLE AND SUBTITLE

Using Data Mining to Determine the Impact Continuity of Care has on the Air Force's Healthcare System

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Wade, David F., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
Wright-Patterson AFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENV-MS-14-D-44

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

711th Human Performance Wing
Colonel Anthony Tvaryanas
Anthony.Tvaryanas@us.af.mil
2610 7th Street, Bldg 441
Wright Patterson AFB, OH 45433-7901
937-255-3814

10. SPONSOR/MONITOR'S ACRONYM(S)

711 HPW/HP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Department of Defense (DoD) healthcare is one of the largest contributors to the DoD budget. In recent years, the cost of the DoD healthcare system has risen at an exponential rate. Much research has been conducted on the impacts that continuity of care has on both improving the quality of patient care and on reducing healthcare costs in the private sector. The DoD has attempted to take a similar approach with regards to healthcare continuity as a means to reduce healthcare costs. This research investigates whether continuity of care influences costs and a military member's availability to perform duties. Specifically, this research examines Air Force fliers with musculoskeletal injuries. Linear and logistic regression techniques are utilized to interpret the relationship continuity of care has on both patient availability and costs. The study does not identify any relationship between continuity of care with costs and patient availability. These findings suggest the need for further research as to whether these findings regarding continuity of care extend beyond musculoskeletal injuries within the DoD healthcare system, as well as evaluating other potential outcomes for continuity of care. Research should also be conducted to determine other factors influencing costs and patient availability.

15. SUBJECT TERMS

Healthcare; Costs; Continuity of Care

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Major Christina Rusnock, AFIT/ENV
U	U	U	UU	87	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636, ext 4611 Christina.Rusnock@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18